LEARNING WITH LESS: MACHINE LEARNING FROM PARTIAL MARKET INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Financial market feeds are sold on the premise that finer granularity and richer reconstruction of hidden structure buy better predictions and better decisions. We test that premise on real Hyperliquid data through three matched studies with one instrument: hold the model, the training budget, the evaluation dates, and the decision policy fixed, then vary only the information the learner sees. The studies ask whether reconstructing hidden order-book structure improves the decision signal, whether forecast-quality gains retain value under executable policies, and whether any fixed book depth, history, or delay is universally sufficient. The answers are consistently negative. On 691 August 2025 breakout events, every hidden summary was reconstructed more accurately than a frozen baseline, yet the reconstruction-augmented predictor was 0.00405 half-Brier worse than the matched coarse predictor and 0.01762 worse than a training-frequency baseline. A multilayer perceptron beat logistic regression on macro-F1 by 0.142 (BTC) and 0.097 (ETH), but every learned policy lost money at 5 basis points per side, and none of the eight contrasts passed the joint gate. And no depth, history, or delay setting produced a common cross-asset benefit across 120 fits. Better intermediate information did not imply better decisions.

1 INTRODUCTION

Electronic markets emit some of the richest sequential data in any applied domain, and a large literature reads that data with modern machine learning: deep networks predict short-horizon price moves from the limit order book (Zhang et al., 2018c; Briola et al., 2024a;b; Sirignano, 2016), order-flow imbalance is engineered into ever finer features (Bechler & Ludkovski, 2014), and execution models chase the last basis point of implementation shortfall (Kato, 2014; Curato et al., 2014). The common working assumption is monotone: a feed with more levels, more history, and a more faithful reconstruction of the hidden book should let a good learner make better predictions and, in turn, better decisions. Data vendors price granularity accordingly.

The assumption is rarely tested as a controlled intervention. Most studies vary the model while holding the information fixed, or improve a forecasting score without asking what the score buys once a decision policy, a spread, a latency, and a cost intervene. When forecast quality and decision quality are examined together, the two can come apart (Falezza, 2026; Raeth & Ludwig, 2025; Gu et al., 2026), and forecast-evaluation tools that sweep decision thresholds already warn that a single accuracy number is a weak surrogate for usefulness (Ziegel et al., 2017; Dimitriadis et al., 2023). What is missing is a clean, matched measurement on real market data: fix everything the learner is allowed to exploit except the information itself, and read the effect at the decision.

We build that instrument and run it three ways on Hyperliquid perpetual-futures data (Figure 1). Each study is a paired comparison in which only the observed information changes. The **reconstruction study** asks whether recovering hidden order-book structure improves the decision signal: we compare a coarse predictor, a predictor augmented with reconstructed hidden summaries, and an observed-rich predictor, all trained and scored identically. The **forecast-to-decision study** asks whether forecast-quality gains survive translation into executable decisions: frozen forecasts from several model families pass through fixed policies under explicit spread, latency, and cost assumptions. The **information-sufficiency factorial** asks whether any fixed book depth, history length, or update delay is universally sufficient: a frozen factorial varies each axis on matched dates. The

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.

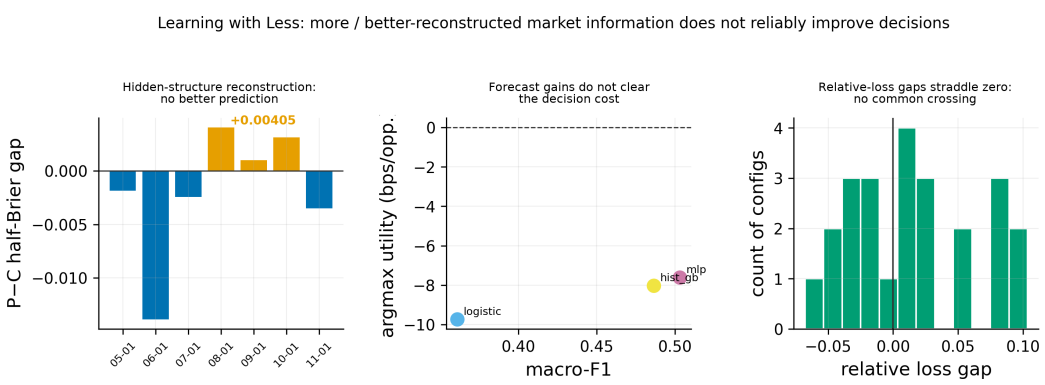

Figure 1: **One instrument, three negative answers.** We hold the model family, training budget, evaluation dates, and decision policy fixed and vary only the information the learner sees, on real Hyperliquid data. **Left (reconstruction):** augmenting a coarse predictor with *more accurately reconstructed* hidden book summaries does not lower prediction loss. On the August breakout cohort it is $+0.00405$ half-Brier *worse* than the coarse predictor. **Center (forecast to decision):** the macro-F1 gain does not clear the cost layer. Every learned policy sits below the zero-utility line at 5 bps per side. **Right (information sufficiency):** the depth/history/delay effect straddles zero across 120 model×asset×factor configurations, with no common crossing.

headline is that all three answers are negative, and they agree: better intermediate information does not reliably become better decisions.

Our contributions are the following.

- **A matched information-intervention protocol** for real order-book data that fixes model, budget, dates, and policy and varies only the observed information, so the effect is read at the decision rather than at a forecasting score (Section 3).
- **A decision-grade negative for reconstruction:** across seven monthly breakout cohorts every hidden summary is reconstructed more accurately than a frozen baseline, yet the reconstruction-augmented predictor is $+0.00405$ half-Brier worse than the coarse predictor and 0.01762 worse than a training-frequency baseline on the 691-event August cohort. Controlled mechanism studies reproduce the same null (Section 4).
- **A forecast-to-decision value audit:** a multilayer perceptron beats logistic regression on macro-F1 by 0.142 (BTC) and 0.097 (ETH) while every learned policy loses money at 5 bps per side. 0 of 8 protocol contrasts pass the joint gate (Section 5).
- **An information-sufficiency factorial:** no depth, history, or delay setting yields a common cross-asset benefit across 120 fits, and the sign of the effect flips with model and asset, so no universal sufficiency threshold exists in this panel (Section 6).
- **A released, laptop-reproducible programme** of thirty curated experiment packages with source certification, provenance, and honest scientific status labels (Section 7). An anonymized copy of the code, results, and this paper is available at `https://anonymous.4open.science/r/learning-with-less-C713`.

The result is not that information is worthless. It is that its value is conditional, small, and easily destroyed by the policy and cost layer that any real decision must cross. We argue this is the more useful thing to measure.

2 Related Work

Deep learning from the limit order book. A large line predicts short-horizon price moves directly from book states, from early convolutional and recurrent models (Sirignano, 2016; Zhang et al., 2018c;b; Nousi et al., 2018; Zheng et al., 2012) through attention and information-persistence

architectures (Jung & Lee, 2024; Briola et al., 2024b; Lucchese et al., 2022) to generative and foundation models of microstructure (Zhang et al., 2026b; Hong et al., 2026; Zhang et al., 2021) and to representation-learning surveys and benchmarks of the book itself (Zhong et al., 2025; Ntakaris et al., 2017; Briola et al., 2024a). This line asks how well a model can predict. It holds the information fixed and varies the model. We invert that: we hold the model fixed and vary the information, and we read the effect at a decision rather than at a forecasting score.

Order-flow structure and queues. A parallel line engineers ever finer views of the book: multi-level and generalized order-flow imbalance (Xu et al., 2019; Cont et al., 2010; Su et al., 2021; Hu & Zhang, 2025; Bechler & Ludkovski, 2014), the dynamics of order positions and queues (Guo et al., 2015; Cont & Larrard, 2012; Toke, 2013; Lipton et al., 2013), and depth forecasting from the book (Libman et al., 2021). These works establish that finer structure carries statistical information. Our reconstruction and information-sufficiency studies ask the next question: whether recovering that structure or paying for more of it changes a decision, and find that on real data it usually does not.

Fill probability, queue position, and execution. Whether an order fills, and at what cost, has been modeled with survival analysis and state-dependent order flow (Arroyo et al., 2023; Lokin & Yu, 2024) and embedded in optimal-execution and market-impact theory (Kato, 2014; 2009; Curato et al., 2014; Said et al., 2018; Brokmann et al., 2014; Lemhadri, 2018; Bechler & Ludkovski, 2017). These give the mechanisms by which book detail *could* matter for a decision. Our finite-queue analysis (Section 4) confirms the mechanism exists but is rare, and our reconstruction diagnostic shows that recovering the hidden queue does not improve the fill/adverse/no-move decision.

Forecast quality versus decision value. The gap between a good forecast and a good decision is documented across domains: rank correlation between accuracy and decision quality breaks down (Falezza, 2026; Raeth & Ludwig, 2025), statistical accuracy is a weak proxy for economic value in electricity and energy markets (Maciejowska et al., 2025; Kath & Ziel, 2018), and decision-focused learning trains to the decision precisely because training to the forecast is insufficient (Wilder et al., 2018; Zhou et al., 2024; Silvestri et al., 2023; Tankov & Tinsi, 2021; Samuel, 2020). Proper-betting guarantees do exist under matched scoring rules and liquidity conditions (Gu et al., 2026). Our forecast-to-decision study supplies the matched market measurement this line calls for: a frozen joint gate on real perpetual-futures data where the forecast winner is not the decision winner.

Forecast evaluation and calibration. Tools that evaluate forecasts across decision thresholds, Murphy diagrams, the Triptych, consistent scoring functions, and calibration diagnostics (Ziegel et al., 2017; Dimitriadis et al., 2023; Ehm et al., 2015; Graziani et al., 2019; Loveday et al., 2023; Taggart, 2021; Allen et al., 2024; Bouallegue et al., 2015), already warn that one accuracy number hides decision-relevant disagreement. We take the warning literally and measure the disagreement in basis points under an explicit policy and cost model.

Benchmarks, leakage, and representation sufficiency. Recent benchmarks probe decision-time leakage and execution realism in learned trading (Zhang et al., 2026a; Yao & Zheng, 2026; Nagy et al., 2025; Cao et al., 2022), backtest correctness and overfitting (Low et al., 2015; Koshiyama & Firoozye, 2019; Pham et al., 2026), and agent-based simulators for counterfactual evaluation (Dwarakanath et al., 2024; Backhouse et al., 2025). On the theory side, decision-sufficient representations preserve the optimal action rather than the input (Ye et al., 2026; Walsh, 2026), and information-theoretic limits bound what any predictor can extract (Catt, 2026). Our released programme operationalizes this stance: matched interventions, frozen gates, source certification, and honest status labels, so a scoped negative is trustworthy rather than an artifact of leakage or a peeked date.

3 THE MATCHED INFORMATION-INTERVENTION INSTRUMENT

Principle. Every study in this paper is one intervention: fix the model family, the training budget, the train/validation/test date split, and the decision policy, then change only the information the learner is allowed to observe. Anything that would let a confound leak in, retuning per view, reusing a peeked date, or scoring a view on its own clock, is held constant by construction. The comparison

is paired at the level of the event, so a difference is attributable to information, not to model capacity or to luck on an easy day.

Data. We use real Hyperliquid perpetual-futures observations. The forecast-to-decision study uses 30 admitted native on-change best-bid/offer contracts across BTC and ETH over 15 paired calendar dates. The information-sufficiency factorial uses 18 admitted sampled-L2 receipt-clock contracts across BTC and ETH over 9 dates. The reconstruction diagnostic uses seven monthly BTC breakout and rebound cohorts drawn from a certified development source. All feature construction, normalization, and target definitions are training-only and frozen before any test date is touched. A separate source-and-schema audit certifies event ordering and timing before an experiment is allowed to run, so a comparison is admitted only when its inputs are valid. Raw market dumps are never redistributed, and each package ships checksums and acquisition instructions instead.

Views and predictors in the reconstruction study. We compare three predictors that differ only in the information they encode. The *coarse* predictor (C) sees an aggregate view. The *augmented* predictor (P) adds machine-reconstructed estimates of hidden book summaries (queue sizes, depth imbalance, and near-touch depth in basis points). And the *observed-rich* predictor (R) sees the true rich summaries. All three share the model family, the training budget, and the dates. We score a multiclass fill/adverse/no-move target with a half-Brier loss (multiclass Brier divided by two), and we anchor every learned arm against two reference baselines: a per-day *uniform* prediction and a *training-frequency* prediction. If reconstruction has decision value, P must beat C and clear the training-frequency reference.

Policies and costs in the forecast-to-decision study. Frozen forecasts from six model families (a prior, logistic regression, gradient boosting, a multilayer perceptron, a random control, and an oracle) pass through three fixed decision policies: an argmax direction policy, a confidence-thresholded policy, and a scheduling policy over 12-opportunity windows. Utility is measured in basis points per opportunity under an explicit execution model with a 100 ms quote offset and a 5 bps-per-side cost. The pre-registered joint gate requires a model to show both a predictive advantage *and* a utility reversal across at least two policy families on six of eight paired dates before a forecast-to-decision improvement is declared.

Information-sufficiency factorial. We cross book depth (one versus five levels), history (instantaneous versus a longer window), and update delay (a current versus a delayed view) and fit each cell with logistic regression, gradient boosting, and a multilayer perceptron. The pre-declared reading is a common crossing: a universal sufficiency threshold would make the same view win across both assets and all model families with simultaneous day-clustered intervals. We report the relative-loss gap of the full depth-plus-history view against the reduced view for every cell.

Soundness. None of the three studies is allowed an outcome-driven tweak: the plans are frozen before test dates are seen, warnings (for example neural fits that reach their iteration cap) are preserved rather than suppressed, and a failure to certify noninferiority is never reported as proof of insufficiency. The evidence we present is a set of scoped negatives and controlled mechanisms, and we label each one by exactly what it does and does not establish.

4 Reconstructing hidden structure does not improve the decision signal

Reconstructing the hidden book more accurately did not make the final prediction better. This is the paper's cleanest result and its most counterintuitive one. A coarse limit-order-book view discards the individual resting orders behind an aggregate volume, and a natural pipeline tries to recover them: estimate the hidden queue sizes and near-touch depth, feed those estimates to the predictor, and expect a lift. We built that pipeline on real BTC breakout and rebound cohorts and measured the lift directly.

Setup recap. On seven monthly cohorts (May through November 2025) we compare the coarse predictor C, the reconstruction-augmented predictor P, and the observed-rich predictor R under the

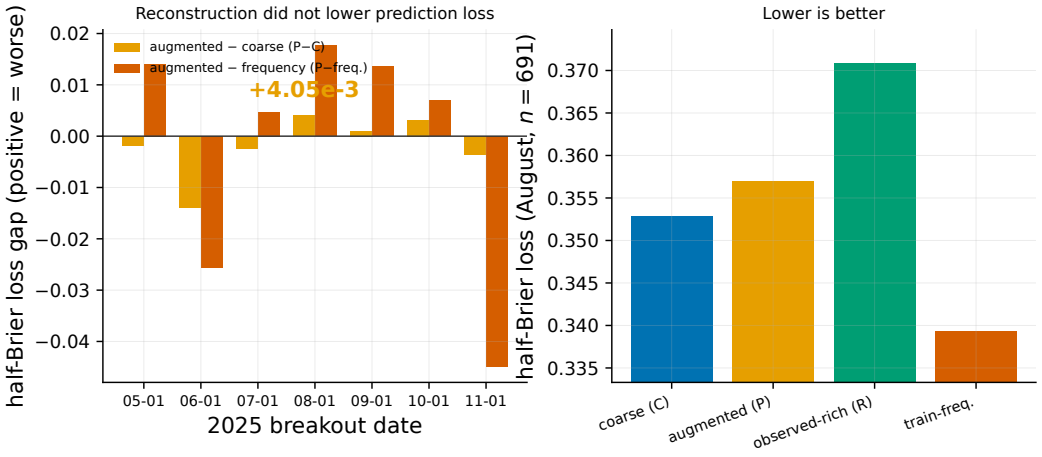

Figure 2: **More accurate reconstruction, no better prediction. Left:** per-date half-Brier loss gap of the augmented predictor P against the coarse predictor C (positive means P is worse) and against the training-frequency baseline. P is worse than the frequency baseline on the majority of dates. The August cohort ($n = 691$) is annotated at $+0.00405$. **Right:** absolute August losses for coarse (C), augmented (P), observed-rich (R), and training-frequency. Lower is better. Neither reconstruction nor the true rich view beats coarse, and none beats the book-free frequency baseline.

identical model, budget, and dates of Section 3, scoring the multiclass target with half-Brier loss. Reconstruction quality is measured separately: every hidden summary that P consumes is compared to a frozen target-mean baseline, and on every cohort each summary is recovered *more* accurately than that baseline. The question is whether that recovered accuracy survives to the decision.

It does not. Figure 2 reads the August cohort, the largest at 691 events. The augmented predictor P scores a half-Brier loss of 0.35692 against the coarse predictor's 0.35286: augmenting with reconstructed structure made the prediction *worse* by 0.00405. It is also 0.01762 worse than the training-frequency baseline, a predictor that ignores the current book entirely and replays the training class mix. The observed-rich predictor R does not rescue the story: it is 0.01793 worse than C on the same cohort. Better recovered inputs, and even the true rich inputs, bought a worse decision signal.

The pattern is not a single unlucky day. Pooling the four August–November breakout dates under equal-day weighting ($n = 1749$), P remains above C ($+0.00118$ half-Brier) and R remains above C ($+0.00318$). No learned breakout arm beats the training-frequency reference under event weighting, and on the earlier development dates the learned arms lose to frequency as well. An earlier aggregate that had appeared to favor rich depth disappeared once models were frozen and re-scored on held-out months: the apparent gain was a class-mix and aggregation artifact, localized by a post-hoc decomposition that separates the class-mixture term from the within-class term.

Controlled mechanism. To check that this is not an estimation failure specific to the market pipeline, we reproduce it in a controlled synthetic setting where the hidden state and the optimal decision are known. There, restricted summaries incur 0.075 decision regret, but a matched direct predictor and the posterior-restoration predictor produce *equal* decision coordinates: hidden-queue uncertainty matters, yet reconstruction carries no isolated advantage over predicting the decision directly, and at 512 training draws the direct and distributional outputs coincide. A conditional shift that is invisible in the observation marginal harms both approaches identically. The mechanism is general: reconstruction spends capacity recovering structure that the decision does not need.

The finite-queue special case. The narrowest reconstruction question, whether knowing the *count* and identity of individual orders narrow which fills are feasible?, has a clean constructed answer that points the same way. In a finite FIFO model over 1000 histories, aggregate volume alone leaves 59 histories ambiguous, and adding displayed order counts leaves 19. Counts narrow the feasible

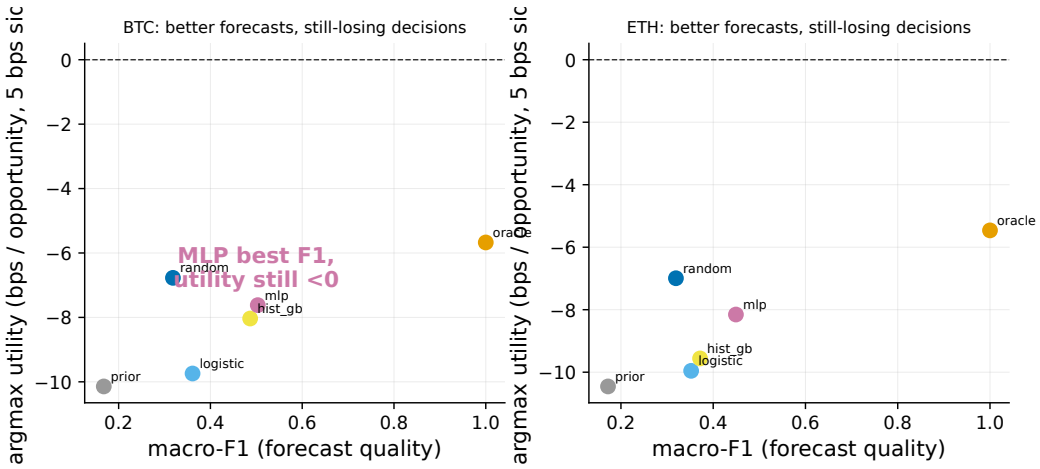

Figure 3: **Better forecasts, still-losing decisions.** Each point is a model family placed at its macro-F1 (x) and its argmax utility in basis points per opportunity at 5 bps per side (y), for BTC (left) and ETH (right). Moving right (better forecasts) does not move above the dashed zero-utility line: every learned policy trades at a loss, and the forecast winner (`mlp`) is not a profitable decision.

outcome in 40 of 1000 histories. The information is real but rare. On admitted native Hyperliquid observations the same count-tightening was not established (0 of 961 in the conditional cohort), and the mandatory native-FIFO endpoint requires an order-lifecycle archive we located but have not yet fully processed (Section 7). The decision-relevance evidence therefore rests on the reconstruction diagnostic above, where the answer is unambiguous: recovering hidden structure did not help the decision.

5 Forecast gains do not establish decision value

The model that forecasts best need not produce a profitable decision. A forecasting score—whether macro-F1, log loss, or direction accuracy—measures a statistical target, but a trader acts through a policy under a spread, a latency, and a cost. The forecast-to-decision study asks whether the gain induced by the score survives translation into that action. We freeze forecasts from six model families and push them through three fixed policies under an explicit execution model (Section 3), on 30 native BBO contracts over 15 paired BTC/ETH dates, completing 22 fits.

The forecast ranking is clear. The multilayer perceptron is the best learned forecaster on both assets: it improves macro-F1 over the logistic baseline by 0.142 on BTC and 0.097 on ETH, with the BTC gain excluding zero on every one of the eight paired dates. On the score alone, one would deploy the perceptron.

The decision ranking erases the win. Figure 3 plots each model at its forecast quality against its argmax utility at 5 bps per side. Every learned policy sits below the zero-utility line: on BTC the logistic policy returns −9.74 bps per opportunity and the perceptron −7.62. On ETH the logistic policy returns −9.96 and the perceptron −8.15. The perceptron loses *less* than logistic regression, so the sign of the forecast advantage is preserved in relative utility, but the absolute decision is to trade at a loss, and the confidence policy's best move is to abstain to exactly zero. The oracle itself, which knows the future label, still returns negative argmax utility (−5.67 BTC, −5.46 ETH) because the utility uses quote endpoints and assumed costs rather than realized fills: even perfect foresight does not clear the cost layer under this policy.

The joint gate fails on every contrast. The pre-registered criterion for declaring a forecast-to-decision improvement requires both a predictive advantage and a utility reversal across at least two policy families on six of eight dates. Of the eight protocol contrasts (one macro-F1 and three utility contrasts per asset), 0 pass. BTC shows a significant macro-F1 gain but the confidence and schedul-

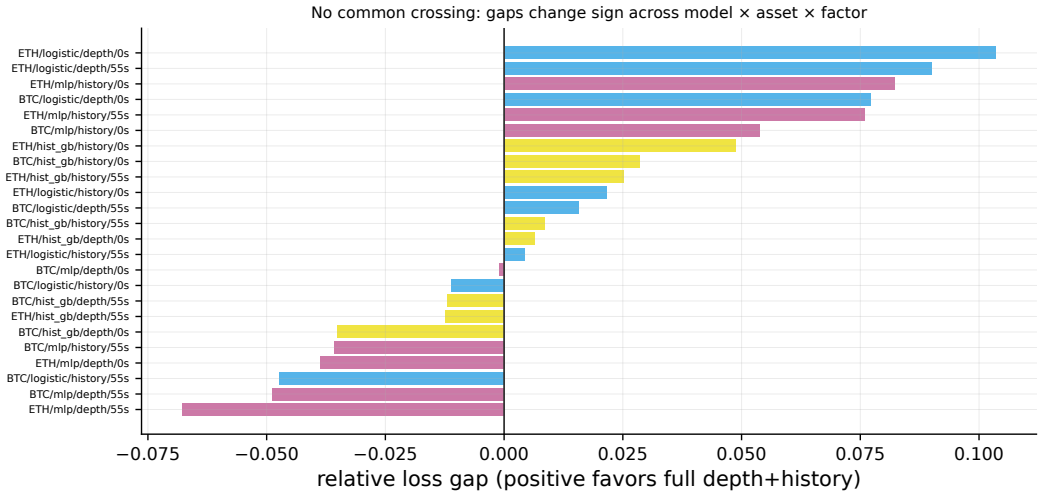

Figure 4: **No common crossing.** Relative-loss gap (full depth+history minus reduced view. Positive favors the richer view) for every model×asset×factor ×delay cell on the transfer dates, sorted. A universal sufficiency threshold would put all bars on one side of zero. Instead the effect changes sign across configurations, and its magnitude is small relative to the spread.

ing contrasts fail their thresholds, and all four ETH contrasts fail. The scheduling policy is negative on four of eight dates for both assets. A forecast advantage that is real on the score is not a decision advantage after the policy and cost layer intervene.

Why this is more than a restatement. That scores and decisions can diverge is known in principle (Falezza, 2026; Raeth & Ludwig, 2025; Ziegel et al., 2017; Dimitriadis et al., 2023), and proper-betting guarantees exist under matched scoring rules and liquidity conditions (Gu et al., 2026). Our contribution is the matched measurement on real perpetual-futures data with a frozen joint gate: the divergence is not a corner case produced by one arbitrary strategy, it is the outcome under the strongest learned forecaster and three standard policies, and it holds on both assets. The forecasting leaderboard and the trading leaderboard are different objects; the matched audit shows how far apart they sit once real costs are paid.

6 No fixed depth, history, or delay is universally sufficient

There is no single view that is enough for every decision. The information-sufficiency factorial tests the vendor's implicit promise most directly: that a deeper, longer, fresher feed is uniformly better, so a buyer can pick one sufficient configuration and stop paying for more. We cross book depth (one versus five levels), history (instantaneous versus a longer window), and update delay (current versus delayed) and fit every cell with three model families on 18 sampled-L2 contracts over 9 BTC/ETH dates, for 120 fits.

The effect has no common sign. Figure 4 shows the relative-loss gap of the full depth-plus-history view against the reduced view for every model×asset×factor cell on the transfer dates. A universal sufficiency threshold would make these gaps line up on one side of zero. They do not. Adding depth helps BTC logistic regression ($+0.077$ relative loss) but hurts the BTC perceptron with a delayed view (-0.049). Adding depth helps ETH logistic regression ($+0.104$) while removing history barely moves ETH gradient boosting. The signs flip with model and with asset, and many simultaneous day-clustered intervals straddle zero, so the panel cannot reject "no effect" for a fixed rule. A separate optimizer diagnostic sharpens the disagreement: shallow inputs fit BTC better while deeper inputs fit ETH better, at both 60 and 600 iterations, so the two assets ask for opposite feeds.

What the negative does and does not say. It does not say depth is useless. Some cells genuinely improve with more information, and the certified reading in a stricter companion analysis (six asset–

model combinations, both assets, at least two families required) certifies a clock-induced change in 0 of 6 cases while explicitly refusing to convert a failure-to-certify-noninferiority into a proof of insufficiency. What the panel rejects is the *simple* rule, "five levels are always enough" or "freshness always makes depth valuable", that a one-size feed decision would need. Extra market detail behaves like adding sensors to a control system: the sensors may carry information, but a particular learner, sample, or timing alignment can fail to exploit it, and here the exploitation fails inconsistently across exactly the axes a buyer would want to hold fixed.

Relation to representation-sufficiency theory. Formal work defines decision-sufficient representations as those that preserve the optimal action (Ye et al., 2026; Walsh, 2026), and depth-aware forecasting studies map predictive value to book levels (Briola et al., 2024b;a). Our panel is the empirical complement: a classifier plateau cannot estimate a theoretical sufficient dimension, but it can show that no fixed observable configuration is sufficient across assets and models on real data. The sufficiency question is therefore conditional by nature, and a leaderboard that reports one average curve hides the sign changes that decide whether a feed is worth buying.

7 SYNTHESIS: INFORMATION VALUE IS CONDITIONAL, SMALL, AND POLICY-FRAGILE

Three questions, asked with one instrument, returned the same answer. Read together, the reconstruction, forecast-to-decision, and information-sufficiency studies rule out the monotone premise that organizes much of applied market-microstructure learning. Reconstructing hidden structure did not improve the decision signal. Forecast gains did not clear the cost layer. No fixed observable configuration was universally sufficient. The common cause is visible once the studies are aligned: value that is real at the level of a statistical target, a recovered queue size, a higher macro-F1, a deeper book, is small in magnitude and is absorbed or reversed by the policy and cost layer that separates a prediction from a decision. This is why we measure at the decision. A programme that had stopped at reconstruction RMSE, at macro-F1, or at an averaged depth curve would have reported three positive results. Measured at the decision, all three are negative.

The reusable object. The contribution readers can depend on is the *matched information-intervention protocol*: freeze model, budget, dates, and policy. Vary only the observed information. Read the effect at the decision under an explicit cost model. And refuse outcome-driven tweaks. The protocol is what turns "does richer data help?" into a falsifiable measurement, and it transfers beyond these three questions to any setting where a feed is bought on the promise of downstream value.

The released programme. We release thirty curated experiment packages spanning the three real-data studies, controlled simulator mechanisms, breakout/rebound cohorts, temporal-persistence checks, and source-and-schema certification. Each package carries its question, method, compact results, figures, provenance, and an explicit scientific-status label, and each is laptop-reproducible from a license-clean smoke fixture. Raw market dumps are never redistributed. The labels are load-bearing: a completed software check, a valid scientific negative, an inconclusive run, and a blocked empirical objective are kept distinct rather than rounded up to "done."

One honest open endpoint. The strictest reconstruction endpoint, native-FIFO count tightening on a fully admitted order-lifecycle cohort, is not yet computed. We located the exact source (an open December 2025 Level-4 Hyperliquid archive of 195.3 GB with 744 hourly files per order archive) and a bounded washout audit over 11 complete hours covered 32,633,820 events, observing 11,328 legacy order identifiers of which 148 were still first seen in hour 10. Observed-lifecycle coverage reached 99.92985%, but true live-book completeness remains unidentified, so we do not claim the native-FIFO endpoint. We flag this rather than paper over it, and the decision-relevance conclusion stands on the reconstruction diagnostic, which needs no such admission.

8 LIMITATIONS

We state the boundaries of the evidence plainly, because a negative result is only as useful as its scope is honest.

Scope of the negatives. The three studies are scoped, not universal. The forecast-to-decision study covers two assets and eight paired dates under a quote-based utility that uses quote endpoints and assumed costs. It does not observe realized fills, endogenous market impact, funding, or queue priority, so it is a decision-utility statement, not a live-trading profitability statement. The information-sufficiency factorial uses sampled-L2 receipt-clock data over nine dates. Its date panel is small and its intervals are wide, so it rejects a simple sufficiency rule rather than proving that depth never helps. The reconstruction diagnostic uses development-exposed dates and is descriptive rather than a pristine prospective holdout.

Data realism. Provider receipt time is uncalibrated against exchange time, and the sampled-L2 feed does not establish native queue priority or calibrated latency. Where an entity or feed could not be certified, the corresponding claim is withheld. The native-FIFO endpoint (Section 7) is the clearest such withholding.

Estimation confounds. Several neural fits reached their fixed iteration budget, and their warnings are preserved rather than suppressed. Full-depth neural underperformance can therefore reflect an optimization limit rather than an information limit. Because changing only that family cannot rescue a result that requires agreement across families, we do not lean on any single neural cell.

What we do not claim. We do not claim that information is worthless, that richer feeds should never be purchased, or that these negatives generalize to other venues, horizons, or asset classes. We claim that, measured at the decision with everything else held fixed, the value of more and better-reconstructed market information was small, conditional, and fragile to the policy and cost layer, across three questions on real data.

9 CONCLUSION

We asked whether more granular and more accurately reconstructed market information buys better decisions, and across three matched studies on real Hyperliquid data the answer was no. Reconstructing hidden book structure left the decision signal no better and often worse, on 691 August breakout events the reconstruction-augmented predictor was 0.00405 half-Brier worse than the coarse predictor and 0.01762 worse than a book-free frequency baseline. Forecast gains did not establish positive decision value: a perceptron won macro-F1 by 0.142 (BTC) and 0.097 (ETH) yet every learned policy lost money at 5 bps per side, and 0 of 8 protocol contrasts passed the joint gate. And no fixed depth, history, or delay was universally sufficient: the effect changed sign across 120 fits.

The through line is that information value lives at the decision, is small there, and is easily undone by the policy and cost layer, so that is where it should be measured. The reusable contribution is the matched information-intervention protocol and a released, laptop-reproducible programme of thirty honestly labeled experiment packages. Learning with less is often not a compromise. On these questions it was indistinguishable from, or better than, learning with more (Ye et al., 2026; Falezza, 2026).

AI USE STATEMENT

Generative AI tools assisted literature processing, code, experimental design, analysis, interpretation, and manuscript drafting and revision. The authors reviewed all AI-assisted work; model-generated reviews and automated checks are not independent human validation. The authors are responsible for the claims, citations, code, and artifacts.

REPRODUCIBILITY STATEMENT

Section 3 specifies the frozen splits, models, policies, and evaluation. The appendices provide per-date results, data constraints, controls, provenance, and threats. The anonymous repository provides thirty reproducible packages with code, results, checksums, smoke fixtures, and acquisition instructions; licensed raw dumps are not redistributed.

REFERENCES

Sam Allen, Jonathan Koh, Johan Segers, and Johanna Ziegel. Tail calibration of probabilistic forecasts. *Journal of the American Statistical Association*, 2024.

Apostolos Ampountolas. Comparative Analysis of Machine Learning, Hybrid, and Deep Learning Forecasting Models Evidence from European Financial Markets and Bitcoins. *Forecasting*, 2023.

Alvaro Arroyo, Alvaro Cartea, Fernando Moreno-Pino, and Stefan Zohren. Deep Attentive Survival Analysis in Limit Order Books: Estimating Fill Probabilities with Convolutional-Transformers. *Quantitative Finance*, 2023.

Alfred Backhouse, Kang Li, Jakob Foerster, Anisoara Calinescu, and Stefan Zohren. Painting the market: generative diffusion models for financial limit order book simulation and forecasting. *arXiv preprint arXiv:2509.05107*, 2025.

Achraf Bahamou, Maud Doumergue, and Philippe Donnat. Hawkes processes for credit indices time series analysis: How random are trades arrival times? *arXiv preprint arXiv:1902.03714*, 2019.

Bastien Baldacci and Philippe Bergault. Optimal incentives in a limit order book: a SPDE control approach. *Market Microstructure and Liquidity*, 2021.

MohammadHossein Bateni, Zahra Hadizadeh, MohammadTaghi Hajiaghayi, Mahdi JafariRaviz, and Shayan Taherijam. Optimal Rates for Agentic Networked Information Aggregation. *arXiv preprint arXiv:2609.05318*, 2026.

Kyle Bechler and Michael Ludkovski. Order Flows and Limit Order Book Resiliency on the Meso-Scale. *Market Microstructure and Liquidity*, 2017.

Kyle Bechler and Mike Ludkovski. Optimal Execution with Dynamic Order Flow Imbalance. *SIAM Journal on Financial Mathematics*, 2014.

Mario Beykirch, Tim Janke, and Florian Steinke. Bidding and Scheduling in Energy Markets: Which Probabilistic Forecast Do We Need? In *2022 17th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, 2022.

Zied Ben Bouallegue, Pierre Pinson, and Petra Friederichs. Quantile forecast discrimination ability and value. *Quarterly Journal of the Royal Meteorological Society*, 2015.

Antonio Briola, Silvia Bartolucci, and Tomaso Aste. Deep Limit Order Book Forecasting. *arXiv preprint arXiv:2403.09267*, 2024a.

Antonio Briola, Silvia Bartolucci, and Tomaso Aste. HLOB – Information Persistence and Structure in Limit Order Books. *Expert Systems with Applications*, 2024b.

X. Brokmann, E. Serie, J. Kockelkoren, and J. P. Bouchaud. Slow decay of impact in equity markets. *Market Microstructure and Liquidity*, 2014.

Defu Cao, Yousef El-Laham, Loc Trinh, Svitlana Vyetrenko, and Yan Liu. DSLOB: A Synthetic Limit Order Book Dataset for Benchmarking Forecasting Algorithms under Distributional Shift. *arXiv preprint arXiv:2211.11513*, 2022.

Tingwei Cao and Yan Xu. An End-to-End Approach for Microgrid Probabilistic Forecasting and Robust Operation via Decision-focused Learning. *arXiv preprint arXiv:2512.12755*, 2025.

Peter Maurice Catt. Forecastability as an Information-Theoretic Limit on Prediction. *arXiv preprint arXiv:2603.27074*, 2026.

Raffaele Giuseppe Cestari, Filippo Barchi, Riccardo Busetto, Daniele Marazzina, and Simone Formentin. Hawkes-based cryptocurrency forecasting via Limit Order Book data. *arXiv preprint arXiv:2312.16190*, 2023.

Umut Cetin and Henri Waelbroeck. Informed trading, limit order book and implementation shortfall: equilibrium and asymptotics. *SSRN Electronic Journal*, 2020.

Gokhan Ceyhan, Nermin Elif Kurt, H. Bahadir Sahin, and Kursad Derinkuyu. Empirical comparison of three models for determining market clearing prices in Turkish day-ahead electricity market. In *2017 14th International Conference on the European Energy Market (EEM)*, 2017.

Jonathan Chavez-Casillas, Jose E. Figueroa-Lopez, Chuyi Yu, and Yi Zhang. Adaptive Optimal Market Making Strategies with Inventory Liquidation Cos. *arXiv preprint arXiv:2405.11444*, 2024.

Jonathan A. Chavez-Casillas and Jose E. Figueroa-Lopez. One-level limit order book models with memory and variable spread. *arXiv preprint arXiv:1407.5684*, 2014.

Shulun Chen, Wei Shao, Flora D. Salim, and Hao Xue. ASTER: Adaptive Spatio-Temporal Early Decision Model for Dynamic Resource Allocation. *arXiv preprint arXiv:2506.17929*, 2025.

Pu Cheng, Juncheng Liu, and Yunshen Long. PolyBench: Benchmarking LLM Forecasting and Trading Capabilities on Live Prediction Market Data. *arXiv preprint arXiv:2604.14199*, 2026.

Rama Cont and Adrien De Larrard. Order book dynamics in liquid markets: limit theorems and diffusion approximations. *arXiv preprint arXiv:1202.6412*, 2012.

Rama Cont and Marvin S. Mueller. A stochastic partial differential equation model for limit order book dynamics. *SIAM Journal on Financial Mathematics*, 2019.

Rama Cont, Arseniy Kukanov, and Sasha Stoikov. The Price Impact of Order Book Events. *Journal of Financial Econometrics*, 2010.

Cameron Cornell, Nam Trong Dinh, and S. Ali Pourmousavi. A probabilistic forecast methodology for volatile electricity prices in the Australian National Electricity Market. *International Journal of Forecasting*, 2023.

Ricardo Crisostomo and Lorena Couso. Financial density forecasts: A comprehensive comparison of risk-neutral and historical schemes. *Journal of Forecasting*, 2018.

Gianbiagio Curato, Jim Gatheral, and Fabrizio Lillo. Optimal execution with nonlinear transient market impact. *Quantitative Finance*, 2014.

Shaw Dalen. What Happens When Institutional Liquidity Enters Prediction Markets: Identification, Measurement, and a Synthetic Proof of Concept. *arXiv preprint arXiv:2604.10005*, 2026.

Timo Dimitriadis, Tilmann Gneiting, Alexander I. Jordan, and Peter Vogel. Evaluating Probabilistic Classifiers: The Triptych. *International Journal of Forecasting*, 2023.

Meik Dorpinghaus, Edgar Roldan, Izaak Neri, Heinrich Meyr, and Frank Julicher. An Information Theoretic Analysis of Sequential Decision-Making. In *2017 IEEE International Symposium on Information Theory (ISIT)*, 2015.

Mouhamad Drame. Limit Order Book (LOB) shape modeling in presence of heterogeneously informed market participants. *Market Microstructure and Liquidity*, 2020.

Philipp D. Dubach. The Anatomy of a Decentralized Prediction Market: Microstructure Evidence from the Polymarket Order Book. *arXiv preprint arXiv:2604.24366*, 2026.

Miroslav Dudik, Rafael Frongillo, and Jennifer Wortman Vaughan. Market Making with Decreasing Utility for Information. *arXiv preprint arXiv:1407.8161*, 2014.

Kshama Dwarakanath, Tucker Balch, and Svitlana Vyetrenko. ABIDES-Economist: Agent-Based Simulator of Economic Systems with Learning Agents. *arXiv preprint arXiv:2402.09563*, 2024.

Werner Ehm, Tilmann Gneiting, Alexander Jordan, and Fabian Kruger. Of Quantiles and Expectiles: Consistent Scoring Functions, Choquet Representations, and Forecast Rankings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2015.

Alessandro Falezza. When Forecast Accuracy Fails: Rank Correlation and Decision Quality in Multi-Market Battery Storage Optimization. *arXiv preprint arXiv:2604.12082*, 2026.

Xinyue Fang and Robert Slepaczuk. Volatility Forecasting and Return Prediction under Market Regimes: Evidence from High-Frequency Chinese Equity Data. *arXiv preprint arXiv:2606.09478*, 2026.

Xuefeng Gao and S. J. Deng. Hydrodynamic limit of order book dynamics. *Probability in the Engineering and Informational Sciences*, 2014.

Lakshya Garg, Sai Yaswanth, Deep Narayan Mishra, Karthik Kumaran, Anupriya Sharma, and Mayank Uniyal. Monodense Deep Neural Model for Determining Item Price Elasticity. In *2026 International Conference on Advances in Artificial Intelligence and Machine Learning (AAIML)*, 2026.

Tim Genewein and Daniel A. Braun. Abstraction in decision-makers with limited information processing capabilities. *arXiv preprint arXiv:1312.4353*, 2013.

Cameron Gordon. The Information Bottleneck Principle in Corporate Hierarchies. *arXiv preprint arXiv:2210.14861*, 2022.

Berend Jelmer Dirk Gort, Xiao-Yang Liu, Xinghang Sun, Jiechao Gao, Shuaiyu Chen, and Christina Dan Wang. Deep Reinforcement Learning for Cryptocurrency Trading: Practical Approach to Address Backtest Overfitting. *arXiv preprint arXiv:2209.05559*, 2022.

Martin D. Gould, Mason A. Porter, and Sam D. Howison. Quasi-Centralized Limit Order Books. *Quantitative Finance*, 2015.

Carlo Graziani, Robert Rosner, Jennifer M. Adams, and Reason L. Machete. Probabilistic Recalibration of Forecasts. *International Journal of Forecasting*, 2019.

Anri Gu, Nicole Kagan, Alec Sun, Jibang Wu, and Haifeng Xu. When do prophets profit in prediction markets? *arXiv preprint arXiv:2607.06166*, 2026.

Xin Guo, Zhao Ruan, and Lingjiong Zhu. Dynamics of Order Positions and Related Queues in a Limit Order Book. *arXiv preprint arXiv:1505.04810*, 2015.

Jiafa He, Cong Zheng, and Can Yang. Integrating Tick-level Data and Periodical Signal for High-frequency Market Making. *arXiv preprint arXiv:2306.17179*, 2023.

Prakul Sunil Hiremath and Vruksha Arun Hiremath. Early Detection of Latent Microstructure Regimes in Limit Order Books. *arXiv preprint arXiv:2604.20949*, 2026.

Simon Hirsch and Florian Ziel. Probabilistic Forecasting for Day-ahead Electricity Prices, Battery Trading Strategies and the Economic Evaluation of Predictive Accuracy. *arXiv preprint arXiv:2604.19580*, 2026.

Xinpeng Hong, Changgang Zheng, Joshua Lilley, Stefan Zohren, and Noa Zilberman. In-Network Market Prediction Using Machine Learning and Limit Order Books. *arXiv preprint arXiv:2608.02424*, 2026.

Chen Hu and Kouxiao Zhang. Stochastic Price Dynamics in Response to Order Flow Imbalance: Evidence from CSI 300 Index Futures. *arXiv preprint arXiv:2505.17388*, 2025.

Weibing Huang, Sergio Pulido, Mathieu Rosenbaum, Pamela Saliba, and Emmanouil Sfendourakis. From Glosten-Milgrom to the whole limit order book and applications to financial regulation. *SSRN Electronic Journal*, 2019.

Ondrej Hubacek and Gustav Sir. Beating the market with a bad predictive model. *International Journal of Forecasting*, 2020.

Jun ichi Maskawa. Correlation of coming limit price with order book in stock markets. *Physica A: Statistical Mechanics and its Applications*, 2007.

Koya Ishikawa and Kazuhide Nakata. Online Trading Models with Deep Reinforcement Learning in the Forex Market Considering Transaction Costs. *arXiv preprint arXiv:2106.03035*, 2021.

Md Khairul Islam, Ayush Karmacharya, Timothy Sue, and Judy Fox. Large Language Models for Financial Aid in Financial Time-series Forecasting. In *2024 IEEE International Conference on Big Data (BigData)*, 2024.

MD Shafikul Islam and Azmine Toushik Wasi. Optimizing Inventory Routing: A Decision-Focused Learning Approach using Neural Networks. *arXiv preprint arXiv:2311.00983*, 2023.

Tim Janke and Florian Steinke. Probabilistic multivariate electricity price forecasting using implicit generative ensemble post-processing. In *2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, 2020.

Johannes Rude Jensen, Mohsen Pourpouneh, Kurt Nielsen, and Omri Ross. The Homogenous Properties of Automated Market Makers. *arXiv preprint arXiv:2105.02782*, 2021.

Jiwon Jung and Kiseop Lee. Attention-Based Reading, Highlighting, and Forecasting of the Limit Order Book. *Quantitative Finance*, 2024.

Christopher Kath and Florian Ziel. The value of forecasts: Quantifying the economic gains of accurate quarter-hourly electricity price forecasts. *Energy Economics*, 2018.

Takashi Kato. An Optimal Execution Problem with Market Impact. *Finance and Stochastics*, 2009.

Takashi Kato. VWAP Execution as an Optimal Strategy. *JSIAM Letters*, 2014.

Adriano Koshiyama and Nick Firoozye. Avoiding Backtesting Overfitting by Covariance-Penalties: an empirical investigation of the ordinary and total least squares cases. *SSRN Electronic Journal*, 2019.

Sebastian M. Krause, Edgar Jungblut, and Thomas Guhr. Two Price Regimes in Limit Order Books: Liquidity Cushion and Fragmented Distant Field. *Journal of Statistical Mechanics: Theory and Experiment*, 2021.

Nermin Elif Kurt, H. Bahadir Sahin, and Kursad Derinkuyu. An Adaptive Tabu Search Algorithm for Market Clearing Problem in Turkish Day-Ahead Market. In *2018 15th International Conference on the European Energy Market (EEM)*, 2018.

Alex Langnau and Yanko Punchev. Stochastic Price Dynamics Implied By the Limit Order Book. *arXiv preprint arXiv:1105.4789*, 2011.

Nam Anh Le. Decomposing Crowd Wisdom: Domain-Specific Calibration Dynamics in Prediction Markets. *arXiv preprint arXiv:2602.19520*, 2026.

Ismael Lemhadri. Market Impact in a Latent Order Book. *arXiv preprint arXiv:1802.06101*, 2018.

Daniel Libman, Simi Haber, and Mary Schaps. Forecasting Quoted Depth With the Limit Order Book. *Frontiers in Artificial Intelligence*, 2021.

Alexander Lipton, Umberto Pesavento, and Michael G Sotiropoulos. Trade arrival dynamics and quote imbalance in a limit order book. *arXiv preprint arXiv:1312.0514*, 2013.

Felix Lokin and Fenghui Yu. Fill Probabilities in a Limit Order Book with State-Dependent Stochastic Order Flows. *arXiv preprint arXiv:2403.02572*, 2024.

Nicholas Loveday, Robert Taggart, and Mohammadreza Khanarmuei. A User-Focused Approach to Evaluating Probabilistic and Categorical Forecasts. *Weather and Forecasting*, 2023.

Robert Low, Stanislaus Maier-Paape, and Andreas Platen. Correctness of Backtest Engines. *The Journal of Investment Strategies*, 2015.

Lorenzo Lucchese, Mikko Pakkanen, and Almut Veraart. The Short-Term Predictability of Returns in Order Book Markets: a Deep Learning Perspective. *International Journal of Forecasting*, 2022.

Katarzyna Maciejowska, Arkadiusz Lipiecki, and Bartosz Uniejewski. Statistical and economic evaluation of forecasts in electricity markets: beyond RMSE and MAE. *Energy Conversion and Management*, 2025.

Stanislaus Maier-Paape and Andreas Platen. Backtest of Trading Systems on Candle Charts. *arXiv preprint arXiv:1412.5558*, 2014.

Vladislav Gennadievich Malyshkin and Ray Bakhramov. Market Dynamics vs. Statistics: Limit Order Book Example. *arXiv preprint arXiv:1603.05313*, 2016.

Peer Nagy, Sascha Frey, Kang Li, Bidipta Sarkar, Svitlana Vyetrenko, Stefan Zohren, Ani Calinescu, and Jakob Foerster. LOB-Bench: Benchmarking Generative AI for Finance – an Application to Limit Order Book Data. *arXiv preprint arXiv:2502.09172*, 2025.

Paraskevi Nousi, Avraam Tsantekidis, Nikolaos Passalis, Adamantios Ntakaris, Juho Kanniainen, Anastasios Tefas, Moncef Gabbouj, and Alexandros Iosifidis. Machine Learning for Forecasting Mid Price Movement using Limit Order Book Data. *arXiv preprint arXiv:1809.07861*, 2018.

Adamantios Ntakaris, Martin Magris, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. Benchmark Dataset for Mid-Price Forecasting of Limit Order Book Data with Machine Learning Methods. *Journal of Forecasting*, 2017.

The Anh Pham, Bao Chan Nguyen, and Nguyet Nguyen Thi. AlgoXpert Alpha Research Framework. A Rigorous IS WFA OOS Protocol for Mitigating Overfitting in Quantitative Strategies. *arXiv preprint arXiv:2603.09219*, 2026.

Elvys Linhares Pontes and Mohamed Benjannet. Contextual Sentence Analysis for the Sentiment Prediction on Financial Data. In *2021 IEEE International Conference on Big Data (Big Data)*, 2021.

Kornelius Raeth and Nicole Ludwig. Forecast Skill Is Not Decision Skill: Evidence from Weather-Dependent Decision Tasks. *arXiv preprint arXiv:2512.14779*, 2025.

Adrian E. Raftery. Use and Communication of Probabilistic Forecasts. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2014.

Jan Rosenzweig. Conservation Laws in a Limit Order Book. *arXiv preprint arXiv:1910.09202*, 2019.

Emilio Said, Ahmed Bel Hadj Ayed, Alexandre Husson, and Frederic Abergel. Market Impact: A Systematic Study of Limit Orders. *Market Microstructure and Liquidity*, 2018.

Jim Samuel. Information Token Driven Machine Learning for Electronic Markets: Performance Effects in Behavioral Financial Big Data Analytics. *Journal of Information Systems and Technology Management*, 2020.

Simon Schnurch and Andreas Wagner. Machine Learning on EPEX Order Books: Insights and Forecasts. *arXiv preprint arXiv:1906.06248*, 2019.

Mattia Silvestri, Senne Berden, Jayanta Mandi, Ali Irfan Mahmutogullar, Brandon Amos, Tias Guns, and Michele Lombardi. Score Function Gradient Estimation to Widen the Applicability of Decision-Focused Learning. *Journal of Artificial Intelligence Research*, 2023.

Justin Sirignano. Deep Learning for Limit Order Books. *Quantitative Finance*, 2016.

Oleg Solozobov. Evidence Sufficiency Under Delayed Ground Truth: Proxy Monitoring for Risk Decision Systems. *arXiv preprint arXiv:2604.15740*, 2026.

Amos Storkey, Jono Millin, and Krzysztof Geras. Isoelastic Agents and Wealth Updates in Machine Learning Markets. *arXiv preprint arXiv:1206.6443*, 2012.

Yuhan Su, Zeyu Sun, Jiarong Li, and Xianghui Yuan. The Price Impact of Generalized Order Flow Imbalance. *arXiv preprint arXiv:2112.02947*, 2021.

Ao Sun and Yuh-Dauh Lyuu. Backtesting Trading Strategies with GAN To Avoid Overfitting. *arXiv preprint arXiv:2209.04895*, 2022.

Robert J. Taggart. Point forecasting and forecast evaluation with generalized Huber loss. *Electronic Journal of Statistics*, 2021.

Peter Tankov and Laura Tinsi. Decision making with dynamic probabilistic forecasts. *arXiv preprint arXiv:2106.16047*, 2021.

Ted Theodosopoulos. Uncertainty relations in models of market microstructure. *Physica A: Statistical Mechanics and its Applications*, 2004.

Ioane Muni Toke. The order book as a queueing system: average depth and influence of the size of limit orders. *Quantitative Finance*, 2013.

Ioane Muni Toke and Nakahiro Yoshida. Modelling intensities of order flows in a limit order book. *Quantitative Finance*, 2016.

Bence Toth and Enrico Scalas. The value of information in financial markets: An agent-based simulation. *arXiv preprint arXiv:0712.2687*, 2007.

Theodoros Tsagaris. Statistical Arbitrage and Optimal Trading with Transaction Costs in Futures Markets. *arXiv preprint arXiv:0801.3348*, 2008.

Svitlana Vyetrenko and Shaojie Xu. Risk-Sensitive Compact Decision Trees for Autonomous Execution in Presence of Simulated Market Response. *arXiv preprint arXiv:1906.02312*, 2019.

Khaled Bin Walid, Feng Ye, Jiaxiang Ji, Ahmed Aziz Ezzat, Travis Miles, and Yazhou Leo Jiang. Economic and Reliability Value of Improved Offshore Wind Forecasting in Bulk Power Grid Operation: A Case Study of The New York Power Grid. *arXiv preprint arXiv:2512.21754*, 2025.

Mark Walsh. Support sufficiency as action-sufficient compression: a single-cycle rate-regret formulation. *arXiv preprint arXiv:2606.09858*, 2026.

Qiuqi Wang, Ruodu Wang, and Johanna Ziegel. E-backtesting. *arXiv preprint arXiv:2209.00991*, 2022.

Yanlong Wang, Jian Xu, Tiantian Gao, Hongkang Zhang, Shao-Lun Huang, Danny Dongning Sun, and Xiao-Ping Zhang. FinTSBridge: A New Evaluation Suite for Real-world Financial Prediction with Advanced Time Series Models. *arXiv preprint arXiv:2503.06928*, 2025.

Edward D. Weinberger. Pragmatic Information Rates, Generalizations of the Kelly Criterion, and Financial Market Efficiency. *arXiv preprint arXiv:0903.2243*, 2009.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the Data-Decisions Pipeline: Decision-Focused Learning for Combinatorial Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Mateusz Wilinski and Juho Kanniainen. Agent-based model of information diffusion in the limit order book trading. *arXiv preprint arXiv:2508.20672*, 2025.

Yao Wu. Limit Order Book Dynamics in Matching Markets: Microstructure, Spread, and Execution Slippage. *arXiv preprint arXiv:2511.20606*, 2025.

Rui Xie, Shuyang Bai, and Ping Ma. Optimal Sampling Designs for Multi-dimensional Streaming Time Series with Application to Power Grid Sensor Data. *The Annals of Applied Statistics*, 2023.

Hai-Chuan Xu, Wei Chen, Xiong Xiong, Wei Zhang, Wei-Xing Zhou, and H Eugene Stanley. Limit-order book resiliency after effective market orders: Spread, depth and intensity. *Journal of Statistical Mechanics: Theory and Experiment*, 2016.

Ke Xu, Martin D. Gould, and Sam D. Howison. Multi-Level Order-Flow Imbalance in a Limit Order Book. *Market Microstructure and Liquidity*, 2019.

Jun Yamada, Karl Pertsch, Anisha Gunjal, and Joseph J. Lim. Task-Induced Representation Learning. *arXiv preprint arXiv:2204.11827*, 2022.

Parley Ruogu Yang. Forecasting high-frequency financial time series: an adaptive learning approach with the order book data. *arXiv preprint arXiv:2103.00264*, 2021.

Junyi Yao and Zihao Zheng. Beyond Agent Architecture: Execution Assumptions and Reproducibility in LLM-Based Trading Systems. *arXiv preprint arXiv:2606.08285*, 2026.

Yuhan Ye, Saurabh Amin, and Asuman Ozdaglar. Learning Decision-Sufficient Representations for Linear Optimization. *arXiv preprint arXiv:2603.18551*, 2026.

Hyung-Jin Yoon and Hunmin Kim. Finite Reliability Representations: Noise-Calibrated Belief-Space Covers for Reliable Decision-Making. *arXiv preprint arXiv:2607.04019*, 2026.

Ali Zahedzadeh and Behnam Bahrak. The Sufficiency-Conciseness Trade-off in LLM Self-Explanation from an Information Bottleneck Perspective. In *Proceedings of the Language Resources and Evaluation Conference*, 2026.

Zhen Zeng, Rachneet Kaur, Suchetha Siddagangappa, Saba Rahimi, Tucker Balch, and Manuela Veloso. Financial Time Series Forecasting using CNN and Transformer. *arXiv preprint arXiv:2304.04912*, 2023.

Hector Zenil and Jean-Paul Delahaye. An algorithmic information-theoretic approach to the behaviour of financial markets. *Journal of Economic Surveys*, 2010.

Fan Zhang, Zhen Li, Sijia Peng, and Yu Chen. When Alpha Disappears: A One-Switch Benchmark for Decision-Time Leakage in Financial Backtests. *arXiv preprint arXiv:2605.23959*, 2026a.

Qiang Zhang, Rui Luo, Yaodong Yang, and Yuanyuan Liu. Benchmarking Deep Sequential Models on Volatility Predictions for Financial Time Series. *arXiv preprint arXiv:1811.03711*, 2018a.

Yanzhi Zhang, Yu Ma, Yilin Cheng, Jian Li, and Yitong Duan. M3: A State-Event Generative Foundation Model for Market Microstructure Dynamics. *arXiv preprint arXiv:2608.19227*, 2026b.

Zihao Zhang, Stefan Zohren, and Stephen Roberts. BDLOB: Bayesian Deep Convolutional Neural Networks for Limit Order Books. *arXiv preprint arXiv:1811.10041*, 2018b.

Zihao Zhang, Stefan Zohren, and Stephen Roberts. DeepLOB: Deep Convolutional Neural Networks for Limit Order Books. *IEEE Transactions on Signal Processing*, 2018c.

Zihao Zhang, Bryan Lim, and Stefan Zohren. Deep Learning for Market by Order Data. *Applied Mathematical Finance*, 2021.

Ban Zheng, Eric Moulines, and Frederic Abergel. Price Jump Prediction in Limit Order Book. *SSRN Electronic Journal*, 2012.

Muyao Zhong, Yushi Lin, and Peng Yang. Representation Learning of Limit Order Book: A Comprehensive Study and Benchmarking. *arXiv preprint arXiv:2505.02139*, 2025.

Hao Zhou, Rongxiao Huang, Shaoming Li, Guibin Jiang, Jiaqi Zheng, Bing Cheng, and Wei Lin. Decision Focused Causal Learning for Direct Counterfactual Marketing Optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.

Johanna F. Ziegel, Fabian Kruger, Alexander Jordan, and Fernando Fasciati. Murphy Diagrams: Forecast Evaluation of Expected Shortfall. *arXiv preprint arXiv:1705.04537*, 2017.

A NOTATION, METRICS, AND DEFINITIONS

Predictors. *Coarse (C)* sees an aggregate order-book view. *Augmented (P)* adds machine-reconstructed estimates of hidden book summaries (queue sizes, depth imbalance, near-touch depth in basis points) to the coarse view. *Observed-rich (R)* sees the true rich summaries. All three share the model family, training budget, and dates. Only the observed information differs.

Targets. The reconstruction study scores a three-class target, fill (F), adverse (A), no-move (N), with a half-Brier loss, defined as the multiclass Brier score divided by two so that a uniform three-class prediction has a fixed reference value. The forecast-to-decision study discretizes a short-horizon mid-quote move into up/down/flat and reports macro-F1. The information-sufficiency factorial reports a relative-loss gap, the fractional change in loss of the full depth-plus-history view against a reduced view.

Baselines. The *uniform* baseline predicts each class with equal probability per day. The *training-frequency* baseline predicts the training class mix and ignores the current book. It is the key reference, because a predictor that cannot beat it has extracted no usable book information. The *oracle* in the forecast-to-decision study has access to the realized label and bounds the best achievable decision utility under the policy and cost model.

Forecast-to-decision policies. The *argmax* policy acts on the highest-probability direction. The *confidence* policy acts only above a fixed probability threshold and otherwise abstains. The *scheduling* policy allocates activity across a 12-opportunity window. Utility is in basis points per opportunity (argmax, confidence) or per window (scheduling), at a 5 bps-per-side cost and a 100 ms quote-offset latency.

Gates. The forecast-to-decision joint gate requires a predictive advantage and a utility reversal across at least two policy families on six of eight paired dates. The information-sufficiency companion requires own-clock inferiority and inherited-clock noninferiority on both assets and at least two model families. Both gates are frozen before test dates are read.

Per-class detail. Table 1 gives the August rebound cohort's per-class breakdown, the companion to Table 3. The same pattern holds, the augmented and rich predictors move the predicted probabilities but not the decision-relevant losses.

Table 1: August rebound cohort ($n = 668$): per-class predicted probability and per-class half-Brier loss by method. Source: `experiments/results/e264_figure_source.json`.

method	q_F	q_A	q_N	ℓ_F	ℓ_A	ℓ_N	total
C	0.089820	0.534431	0.375749	0.727097	0.282014	0.251431	0.310500
P	0.089820	0.534431	0.375749	0.727468	0.289412	0.239786	0.310111
R	0.089820	0.534431	0.375749	0.759518	0.331935	0.234081	0.333572
onehot-F	0.089820	0.534431	0.375749	0.000000	1.000000	1.000000	0.910180
onehot-A	0.089820	0.534431	0.375749	1.000000	0.000000	1.000000	0.465569
onehot-N	0.089820	0.534431	0.375749	1.000000	1.000000	0.000000	0.624251
uniform	0.089820	0.534431	0.375749	0.333333	0.333333	0.333333	0.333333
train-frequency	0.089820	0.534431	0.375749	0.643522	0.254912	0.253237	0.289188

B RECONSTRUCTION STUDY: PROTOCOL AND FULL RESULTS

This appendix expands Section 4. The reconstruction study compares three predictors that differ only in the information they encode, coarse (C), reconstruction augmented (P), and observed rich (R), against two reference baselines (a per-day uniform prediction and a training-frequency prediction), scoring a three-class fill/adverse/no-move target with half-Brier loss (multiclass Brier divided by two). All arms share the model family, the training budget, and the calendar dates. Only the observed information differs.

Per-date losses. Table 2 lists every predictor's half-Brier loss for each monthly cohort, together with the two decision-relevant gaps: P−C (does augmenting with reconstructed structure beat the coarse view?) and P−frequency (does it beat a book-free baseline?). On the August breakout cohort ($n = 691$) the augmented predictor P scores a half-Brier loss of 0.35692, against the coarse predictor's 0.35286 and the training-frequency baseline's 0.33930. The gap to coarse is $+0.00405$ and the gap to frequency is $+0.01762$: the augmented predictor is worse on both comparisons. The pattern is not confined to August. Across the breakout cohorts the augmented predictor fails to beat the training-frequency baseline on the majority of dates, and the observed-rich predictor R does no better, confirming that the failure is about the decision target rather than about reconstruction error.

Table 2: Reconstruction study: per-date half-Brier loss by predictor, with the two decision-relevant gaps. C = coarse, P = reconstruction-augmented, R = observed-rich. Positive gaps mean the augmented predictor is worse. All values from `experiments/results/e264_figure_source.json`.

family	date	n	C	P	R	uniform	train-freq.	P−C	P−freq.
breakout	2025-05-01	372	0.310878	0.309019	0.302571	0.333333	0.294974	-0.001858	0.014046
breakout	2025-06-01	155	0.235191	0.221313	0.211069	0.333333	0.247009	-0.013879	-0.025697
breakout	2025-07-01	247	0.288472	0.286019	0.276986	0.333333	0.281304	-0.002453	0.004715
breakout	2025-08-01	691	0.352863	0.356918	0.370796	0.333333	0.339301	0.004054	0.017616
breakout	2025-09-01	594	0.299716	0.300729	0.302344	0.333333	0.287154	0.001013	0.013575
breakout	2025-10-01	360	0.283368	0.286514	0.279983	0.333333	0.279556	0.003145	0.006958
breakout	2025-11-01	104	0.203500	0.199986	0.199041	0.333333	0.244839	-0.003514	-0.044853
rebound	2025-05-01	317	0.300703	0.293038	0.303007	0.333333	0.283472	-0.007665	0.009566
rebound	2025-06-01	122	0.273435	0.272781	0.273255	0.333333	0.260157	-0.000654	0.012624
rebound	2025-07-01	195	0.281781	0.285348	0.290173	0.333333	0.267969	0.003567	0.017379
rebound	2025-08-01	668	0.310500	0.310111	0.333572	0.333333	0.289188	-0.000389	0.020923
rebound	2025-09-01	574	0.304925	0.304402	0.298784	0.333333	0.279154	-0.000523	0.025248
rebound	2025-10-01	311	0.301289	0.300011	0.314831	0.333333	0.276532	-0.001277	0.023480
rebound	2025-11-01	80	0.259826	0.252390	0.280369	0.333333	0.263518	-0.007436	-0.011127

Decomposition. A post-hoc decomposition separates each loss gap into a class-mixture term (differences in the realized class frequencies across cohorts) and a within-class term (differences in per-class loss). The earlier aggregate that had appeared to favor rich depth is localized to the class-mixture term and to equal-versus-pooled day weighting. Under event weighting, no learned breakout arm beats the training-frequency reference. This is why we report the gap under a fixed weighting and against an explicit baseline rather than as a single aggregate number: the aggregate is exactly where the artifact hid.

Per-class breakdown of the headline cohort. Table 3 opens the August cohort by class. The augmented predictor P and the observed-rich predictor R shift the predicted class probabilities toward the recovered structure, but the per-class losses do not improve over the coarse predictor C on the classes that matter for the decision. The total loss rises accordingly. The one-hot rows confirm that the target is genuinely three-way and that no single class dominates the loss, so the null is not an artifact of a degenerate label.

Table 3: August breakout cohort ($n = 691$): per-class predicted probability (q) and per-class half-Brier loss (ℓ) for each method, with total loss. Classes are fill (F), adverse (A), and no-move (N). Source: `experiments/results/e264_figure_source.json`.

method	q_F	q_A	q_N	ℓ_F	ℓ_A	ℓ_N	total
C	0.367583	0.231548	0.400868	0.375428	0.680592	0.142871	0.352863
P	0.367583	0.231548	0.400868	0.400633	0.688576	0.125261	0.356918
R	0.367583	0.231548	0.400868	0.440539	0.708046	0.112043	0.370796
onehot-F	0.367583	0.231548	0.400868	0.000000	1.000000	1.000000	0.632417
onehot-A	0.367583	0.231548	0.400868	1.000000	0.000000	1.000000	0.768452
onehot-N	0.367583	0.231548	0.400868	1.000000	1.000000	0.000000	0.599132
uniform	0.367583	0.231548	0.400868	0.333333	0.333333	0.333333	0.333333
train-frequency	0.367583	0.231548	0.400868	0.366264	0.578116	0.176634	0.339301

Controlled mechanism. The synthetic control of Section 4 places the same question in a setting with a known hidden state and a known optimal action. Restricted summaries incur 0.075 decision regret relative to the full-information optimum, but a matched direct predictor and a posterior-restoration predictor produce equal decision coordinates, and at 512 training draws their outputs coincide. The lesson transfers: when the decision does not require the hidden structure, reconstructing it spends capacity without changing the action, and when a conditional shift is invisible in the observation marginal it harms both approaches equally. Reconstruction is therefore neither a free lunch nor a targeted fix. It is orthogonal to the decision unless the decision needs precisely the recovered coordinate.

Finite-queue special case. The constructed finite FIFO analysis isolates the narrowest version of the reconstruction question: can the displayed *count* of resting orders narrow which fills are feasible beyond what aggregate volume already implies? Over 1000 enumerated histories, aggregate volume leaves 59 ambiguous and adding counts leaves 19. Counts strictly narrow the feasible outcome in 40 histories. The tightening is real but rare, and on the admitted native Hyperliquid cohort the same tightening was not observed (0 of 961 primary. 0 of 962 in a postrun completion check), consistent with the reconstruction diagnostic: the extra information exists but seldom changes the decision.

C FORECAST-TO-DECISION STUDY: PROTOCOL AND FULL RESULTS

Data and features. The study uses 30 admitted native on-change best-bid/offer contracts across BTC and ETH over 15 paired calendar dates. Features are past-only quote statistics computed on the receipt clock. The target is the sign of a short-horizon mid-quote move discretized into up/down/flat. Feature construction and normalization are fit on the training portion only and frozen before any test date is scored. Six model families produce frozen forecasts: a class-prior baseline, logistic regression, gradient boosting, a multilayer perceptron, a random control, and an oracle with access to the realized label.

Policies and utility. Each frozen forecast passes through three fixed policies. The *argmax* policy takes the highest-probability direction. The *confidence* policy acts only above a fixed probability threshold and otherwise abstains. The *scheduling* policy allocates activity across a 12-opportunity window. Utility is computed with inverse-contract price arithmetic and scenario fees at a 5 bps-per-side cost and a 100 ms quote-offset latency. Utility uses quote endpoints and assumed costs. It does not observe realized fills, funding, or endogenous impact, so it is a decision utility, not a realized profit.

Joint gate. A forecast-to-decision improvement is declared only when a model shows both a predictive advantage and a utility reversal across at least two policy families on six of eight paired dates. Table 4 reports the four primary contrasts per asset (one macro-F1 and three utility contrasts). No contrast passes: the BTC macro-F1 gain is significant, but the confidence and scheduling contrasts fail, and all four ETH contrasts fail. The confidence policy collapses to zero utility (abstain) for every learned model, and the scheduling contrast is negative on four of eight dates for both assets.

Table 4: Forecast-to-decision primary contrasts (MLP minus logistic) with bootstrap intervals over eight paired dates. Source: `experiments/results/q17_primary_contrasts.csv`. No contrast passes the joint gate.

asset	metric	mean	95% interval	neg. dates	passed
BTC	macro_f1	0.142103	[0.039529, 0.244677]	0	✓
BTC	argmax	2.122375	[0.292744, 3.952005]	0	×
BTC	confidence	0.000000	[0.000000, 0.000000]	0	×
BTC	scheduling	0.025961	[-0.102278, 0.154200]	4	×
ETH	macro_f1	0.097081	[0.006888, 0.187274]	1	×
ETH	argmax	1.801718	[0.096671, 3.506765]	1	×
ETH	confidence	0.000000	[0.000000, 0.000000]	0	×
ETH	scheduling	-0.070943	[-0.524662, 0.382776]	4	×

Table 5: Forecast-to-decision absolute performance by model family. Utility in basis points per opportunity (argmax, confidence) or per 12-opportunity window (scheduling) at 5 bps per side. Source: `experiments/results/q17_figure_data.json`.

asset	model	macro-F1	argmax	confidence	scheduling
BTC	prior	0.167633	-10.143430	0.000000	-0.087832
BTC	logistic	0.361085	-9.741636	0.000000	0.680911
BTC	hist_gb	0.486577	-8.034500	0.000000	0.531820
BTC	mlp	0.503188	-7.619262	0.000000	0.706872
BTC	random	0.318313	-6.772006	0.000000	-0.080510
BTC	oracle	1.000000	-5.673321	0.000000	1.997775
ETH	prior	0.171960	-10.455921	0.000000	-0.108507
ETH	logistic	0.352168	-9.956131	0.000000	0.903621
ETH	hist_gb	0.371520	-9.559819	0.000000	0.797432
ETH	mlp	0.449249	-8.154412	0.000000	0.832678
ETH	random	0.319194	-6.992946	0.000000	-0.080479
ETH	oracle	1.000000	-5.461902	0.000000	2.752498

Table 6: Forecast-to-decision study: per-date macro-F1 advantage of the MLP over logistic regression, by paired date, with the mean. Source: `experiments/results/q17_figure_data.json`.

asset	d1	d2	d3	d4	d5	d6	d7	d8	mean
BTC	0.223477	0.072209	0.265424	0.076891	0.089394	0.088852	0.194926	0.125649	**0.142103**
ETH	-0.001203	0.073973	0.190915	0.060807	0.086582	0.080443	0.196469	0.088661	**0.097081**

Table 6 gives the per-date macro-F1 advantage: the BTC gain is positive on all eight dates (hence the significant contrast), while the ETH gain is negative on one date, which is why the ETH predictive contrast is weaker even though its mean is positive.

Absolute performance. Table 5 gives the absolute macro-F1 and per-policy utility for every model family. The ordering by macro-F1 (oracle $>$ mlp $>$ hist_gb $>$ logistic $>$ random $>$ prior) does not match the ordering by argmax utility, and every learned argmax utility is negative. The oracle's own argmax utility is negative (-5.67 BTC, -5.46 ETH), which bounds how much the cost layer alone removes: even a perfect direction forecast trades at a loss under this policy and cost model. This is the quantitative core of the section: the forecasting leaderboard and the decision leaderboard are different orderings, and the gap is large enough to invert the deployment choice.

Sensitivity across later dates. A frozen follow-up sweeps fixed cost and latency scenarios on additional dates without changing the gate. The per-policy utility ranges remain policy-dependent and continue to straddle zero for the confidence and scheduling families, so the negative is not an artifact of one cost assumption. We report ranges rather than a single tuned scenario precisely to avoid selecting the cost point that would flip the conclusion.

D Information-sufficiency factorial: protocol and full results

Design. The factorial crosses three information axes, book depth (one versus five levels), history (instantaneous versus a longer window), and update delay (a current view versus a view delayed by 55 seconds), and fits each cell with logistic regression, gradient boosting, and a multilayer perceptron on 18 admitted sampled-L2 receipt-clock contracts across BTC and ETH over 9 dates, for 120 fits. The response is the relative-loss gap of the full depth-plus-history view against the reduced view. A positive gap favors the richer view. Normalization and feature transforms are training-only.

No common crossing. Table 7 lists every cell for both the pilot and transfer stages. A universal sufficiency threshold would place all transfer gaps on one side of zero with simultaneous intervals excluding it. Instead the gaps change sign across model, asset, and factor: on transfer dates BTC

logistic regression gains $+0.077$ from depth while the BTC perceptron with a delayed view loses -0.049. ETH logistic regression gains $+0.104$ from depth while ETH gradient boosting with a delayed view is slightly negative. Many simultaneous upper bounds sit on the opposite side of the point estimate, so the panel cannot reject "no effect" for a fixed rule.

Table 7: Information-sufficiency factorial: relative-loss gap of the full depth+history view against the reduced view (positive favors the richer view), with simultaneous upper bounds on transfer dates. Source: `experiments/results/q18_contrasts.csv`.

stage	asset	model	delay	factor	rel. loss	sim. upper
pilot	BTC	logistic	0	depth	0.009179	–
pilot	BTC	logistic	0	history	0.020433	–
pilot	BTC	logistic	55	depth	-0.006219	–
pilot	BTC	logistic	55	history	0.012249	–
pilot	BTC	hist_gb	0	depth	0.015945	–
pilot	BTC	hist_gb	0	history	0.022297	–
pilot	BTC	hist_gb	55	depth	0.020958	–
pilot	BTC	hist_gb	55	history	0.019063	–
pilot	ETH	logistic	0	depth	-0.003899	–
pilot	ETH	logistic	0	history	-0.011585	–
pilot	ETH	logistic	55	depth	-0.056422	–
pilot	ETH	logistic	55	history	-0.016132	–
pilot	ETH	hist_gb	0	depth	0.078818	–
pilot	ETH	hist_gb	0	history	0.021399	–
pilot	ETH	hist_gb	55	depth	0.008940	–
pilot	ETH	hist_gb	55	history	0.019923	–
transfer	BTC	logistic	0	depth	0.077166	0.239233
transfer	BTC	logistic	0	history	-0.011160	0.043219
transfer	BTC	logistic	55	depth	0.015667	0.070315
transfer	BTC	logistic	55	history	-0.047349	0.013859
transfer	BTC	hist_gb	0	depth	-0.035137	0.014592
transfer	BTC	hist_gb	0	history	0.028600	0.035509
transfer	BTC	hist_gb	55	depth	-0.011948	0.016189
transfer	BTC	hist_gb	55	history	0.008541	0.011751
transfer	BTC	mlp	0	depth	-0.001089	0.083898
transfer	BTC	mlp	0	history	0.053874	0.140525
transfer	BTC	mlp	55	depth	-0.048831	-0.026603
transfer	BTC	mlp	55	history	-0.035745	0.071483
transfer	ETH	logistic	0	depth	0.103533	0.303054
transfer	ETH	logistic	0	history	0.021481	0.080885
transfer	ETH	logistic	55	depth	0.089966	0.218659
transfer	ETH	logistic	55	history	0.004337	0.084905
transfer	ETH	hist_gb	0	depth	0.006368	0.112091
transfer	ETH	hist_gb	0	history	0.048809	0.068430
transfer	ETH	hist_gb	55	depth	-0.012304	-0.008693
transfer	ETH	hist_gb	55	history	0.025069	0.048440
transfer	ETH	mlp	0	depth	-0.038786	0.203815
transfer	ETH	mlp	0	history	0.082120	0.134936
transfer	ETH	mlp	55	depth	-0.067779	0.148440
transfer	ETH	mlp	55	history	0.075997	0.140244

Optimizer diagnostic. A separate check varies the neural optimizer budget to rule out a pure underfitting explanation. Shallow inputs fit BTC better while deeper inputs fit ETH better, at both 60 and 600 iterations. The asset-dependent direction survives the budget change, so it is not merely an artifact of an unconverged fit. Several neural cells did reach their iteration cap, and their warnings are preserved. Because a universal rule would require agreement across model families, no single neural cell can establish or rescue the result.

Table 8: Information-sufficiency optimizer diagnostic: own-clock loss at depth 1 and depth 5 by asset and neural budget. A positive shallow-minus-deep value means shallow inputs fit better. Source: `experiments/results/q18_figure_data.json`.

asset	budget	depth-1 loss	depth-5 loss	shallow − deep
BTC	60	0.307662	0.314045	-0.006383
BTC	600	0.307170	0.316755	-0.009585
ETH	60	0.291022	0.277982	0.013041
ETH	600	0.289594	0.281961	0.007633

Table 8 gives the optimizer diagnostic in full: shallow inputs fit BTC better (negative shallow-minus-deep) while deeper inputs fit ETH better (positive), at both budgets. Table 9 shows the per-date components of the transfer-stage gaps, and makes the instability concrete: within a single cell the daily gap changes sign across dates, so the cell-level average is an average over disagreement rather than a stable effect.

Table 9: Information-sufficiency per-date relative-loss gaps (transfer stage): the first three daily components of each cell, showing how the sign varies day to day. Source: `experiments/results/q18_figure_data.json`.

asset	model	delay	factor	day 1	day 2	day 3
BTC	logistic	0	depth	-0.029221	0.021485	0.239233
BTC	logistic	0	history	0.043219	0.006853	-0.083553
BTC	logistic	55	depth	-0.012472	-0.010842	0.070315
BTC	logistic	55	history	0.013859	-0.031358	-0.124548
BTC	hist_gb	0	depth	-0.064672	-0.055331	0.014592
BTC	hist_gb	0	history	0.029717	0.035509	0.020573
BTC	hist_gb	55	depth	-0.031816	-0.020217	0.016189
BTC	hist_gb	55	history	0.008076	0.005796	0.011751
BTC	mlp	0	depth	-0.060982	-0.026183	0.083898
BTC	mlp	0	history	0.140525	0.124402	-0.103304
BTC	mlp	55	depth	-0.060315	-0.059575	-0.026603
BTC	mlp	55	history	0.071483	0.017765	-0.196483
ETH	logistic	0	depth	-0.076023	0.083567	0.303054
ETH	logistic	0	history	0.080885	0.054798	-0.071239
ETH	logistic	55	depth	-0.052137	0.103377	0.218659
ETH	logistic	55	history	0.084905	0.037253	-0.109145
ETH	hist_gb	0	depth	-0.062075	-0.030913	0.112091
ETH	hist_gb	0	history	0.068430	0.060063	0.017933
ETH	hist_gb	55	depth	-0.014729	-0.013490	-0.008693
ETH	hist_gb	55	history	0.048440	0.028426	-0.001659
ETH	mlp	0	depth	-0.192436	-0.127737	0.203815
ETH	mlp	0	history	0.132405	0.134936	-0.020980
ETH	mlp	55	depth	-0.198363	-0.153413	0.148440
ETH	mlp	55	history	0.140244	0.121247	-0.033501

Certified companion. A stricter companion analysis on an independent depth panel requires an own-clock inferiority *and* an inherited-clock noninferiority, on both assets and at least two model families, before certifying a clock-induced sufficiency change. It certifies the change in 0 of 6 asset–model cases and explicitly refuses to convert a failure-to-certify-noninferiority into a proof of insufficiency. The two analyses agree on the conclusion that matters here: no fixed observable configuration is universally sufficient, and the sufficiency question is conditional on model, asset, and timing.

E EXTENDED RESULTS AND PROTOCOL

All reconstruction contrasts. Table 10 reports every pairwise loss contrast in the reconstruction study, across the breakout and rebound families, the May–July and August–November cohort groups, and both the equal-day and pooled weightings, in raw and standardized form. The pattern that the main text summarizes is visible throughout the table. The augmented-minus-coarse (P−C) and rich-minus-coarse (R−C) contrasts are small and frequently positive, so augmentation and rich inputs lose. The predictor-minus-frequency contrasts show that no learned arm reliably beats the book-free baseline. The frequency-minus-uniform contrast confirms that the training-frequency baseline is itself a strong reference. The rich-minus-augmented (R−P) contrast is small in every cell. This is why reconstruction quality does not translate into a decision: the observed-rich predictor and its reconstruction do not separate at the decision even where the reconstruction is accurate.

Table 10: Reconstruction study: all pairwise loss contrasts by family, cohort group, and weighting (raw and standardized). Positive means the first predictor loses more. Source: `experiments/results/e264_group_metrics.csv`.

family	group	weight	contrast	n	raw	standardized

Forecast-to-decision cost/latency sensitivity. Table 11 reports, for each asset and policy, the utility range swept across fixed later-date cost and latency scenarios, together with the number of cells whose sign opposes the point estimate. The argmax policy is consistently negative in absolute terms, the confidence policy collapses to abstention, and the scheduling policy straddles zero with many opposing cells. The negative therefore does not depend on a single chosen cost point. It is stable across the scenario sweep, which is the reason we report ranges rather than a tuned scenario.

Table 11: Forecast-to-decision later-date sensitivity: utility range across fixed cost/latency scenarios per asset and policy, with the count of cells whose sign opposes the point estimate. Source: `experiments/results/q17_figure_data.json`.

asset	policy	min	max	opposite	cells	macro-F1
BTC	argmax	0.0017	4.3955	0	54	0.1799
BTC	fixed_confidence	-1.7188	0.0577	45	54	0.1799
ETH	argmax	0.0279	5.2932	0	54	0.1585
ETH	fixed_confidence	-0.7974	0.0039	48	54	0.1585

Mixture-versus-within decomposition. Table 12 decomposes each reconstruction loss gap into a class-mixture term (driven by differences in realized class frequencies across cohorts) and a within-class term (driven by per-class loss differences). The mixture term carries most of any apparent movement, which is exactly why an aggregate over heterogeneous cohorts can manufacture a depth "advantage" that vanishes under a fixed weighting: the aggregate is confounded by the changing class mix, not by information. The within-class terms, which isolate genuine predictive change, are small and do not favor the augmented or rich predictors.

Table 12: Reconstruction loss-gap decomposition into a class-mixture term and a within-class term, by family, contrast, and metric. The mixture term localizes the earlier apparent depth advantage. Source: `experiments/results/e264_decompositions.csv`.

family	contrast	metric	status	Δ	mix	within
breakout	equal_day	C	evaluable	0.006682	0.020763	-0.014081
breakout	equal_day	P	evaluable	0.013920	0.021523	-0.007603
breakout	equal_day	R	evaluable	0.024499	0.022805	0.001694
breakout	equal_day	onehot-F	evaluable	-0.021897	-0.021897	0.000000
breakout	equal_day	onehot-A	evaluable	-0.022745	-0.022745	0.000000

family	contrast	metric	status	Δ	mix	within
breakout	equal_day	onehot-N	evaluable	0.044642	0.044642	0.000000
breakout	equal_day	uniform	evaluable	-0.000000	0.000000	-0.000000
breakout	equal_day	train-frequency	evaluable	0.013284	0.013284	0.000000
breakout	equal_day	P-C	evaluable	0.007238	0.000760	0.006478
breakout	equal_day	R-C	evaluable	0.017817	0.002042	0.015775
breakout	equal_day	R-P	evaluable	0.010579	0.001282	0.009297
breakout	equal_day	C-frequency	evaluable	-0.006602	0.007479	-0.014081
breakout	equal_day	P-frequency	evaluable	0.000636	0.008239	-0.007603
breakout	equal_day	R-frequency	evaluable	0.011215	0.009521	0.001694
breakout	equal_day	frequency-uniform	evaluable	0.013284	0.013284	0.000000
breakout	pooled	C	evaluable	0.023057	0.035486	-0.012429
breakout	pooled	P	evaluable	0.029897	0.036989	-0.007092
breakout	pooled	R	evaluable	0.042561	0.039458	0.003103
breakout	pooled	onehot-F	evaluable	-0.053973	-0.053973	0.000000
breakout	pooled	onehot-A	evaluable	-0.030975	-0.030975	0.000000
breakout	pooled	onehot-N	evaluable	0.084948	0.084948	0.000000
breakout	pooled	uniform	evaluable	-0.000000	0.000000	0.000000
breakout	pooled	train-frequency	evaluable	0.022671	0.022671	-0.000000
breakout	pooled	P-C	evaluable	0.006839	0.001503	0.005337
breakout	pooled	R-C	evaluable	0.019504	0.003972	0.015532
breakout	pooled	R-P	evaluable	0.012664	0.002469	0.010195
breakout	pooled	C-frequency	evaluable	0.000386	0.012815	-0.012429
breakout	pooled	P-frequency	evaluable	0.007226	0.014318	-0.007092
breakout	pooled	R-frequency	evaluable	0.019890	0.016787	0.003103
breakout	pooled	frequency-uniform	evaluable	0.022671	0.022671	0.000000
rebound	equal_day	C	evaluable	0.008829	0.008453	0.000375
rebound	equal_day	P	evaluable	0.008006	0.009437	-0.001431
rebound	equal_day	R	evaluable	0.018077	0.014875	0.003202
rebound	equal_day	onehot-F	evaluable	-0.016623	-0.016623	0.000000
rebound	equal_day	onehot-A	evaluable	-0.046166	-0.046166	0.000000
rebound	equal_day	onehot-N	evaluable	0.062788	0.062788	0.000000
rebound	equal_day	uniform	evaluable	0.000000	-0.000000	-0.000000
rebound	equal_day	train-frequency	evaluable	0.006565	0.006565	-0.000000
rebound	equal_day	P-C	evaluable	-0.000823	0.000984	-0.001806
rebound	equal_day	R-C	evaluable	0.009249	0.006422	0.002827
rebound	equal_day	R-P	evaluable	0.010071	0.005438	0.004633
rebound	equal_day	C-frequency	evaluable	0.002264	0.001888	0.000375
rebound	equal_day	P-frequency	evaluable	0.001441	0.002872	-0.001431
rebound	equal_day	R-frequency	evaluable	0.011512	0.008310	0.003202
rebound	equal_day	frequency-uniform	evaluable	0.006565	0.006565	-0.000000
rebound	pooled	C	evaluable	0.014667	0.008656	0.006011
rebound	pooled	P	evaluable	0.016578	0.010108	0.006470
rebound	pooled	R	evaluable	0.021834	0.017754	0.004080
rebound	pooled	onehot-F	evaluable	-0.019597	-0.019597	0.000000
rebound	pooled	onehot-A	evaluable	-0.075886	-0.075886	0.000000
rebound	pooled	onehot-N	evaluable	0.095482	0.095482	0.000000
rebound	pooled	uniform	evaluable	-0.000000	0.000000	0.000000
rebound	pooled	train-frequency	evaluable	0.007775	0.007775	-0.000000
rebound	pooled	P-C	evaluable	0.001911	0.001452	0.000458
rebound	pooled	R-C	evaluable	0.007166	0.009098	-0.001931
rebound	pooled	R-P	evaluable	0.005256	0.007646	-0.002390
rebound	pooled	C-frequency	evaluable	0.006892	0.000881	0.006011
rebound	pooled	P-frequency	evaluable	0.008803	0.002333	0.006470
rebound	pooled	R-frequency	evaluable	0.014058	0.009979	0.004080
rebound	pooled	frequency-uniform	evaluable	0.007775	0.007775	-0.000000

Protocol in pseudocode. The matched intervention is deliberately simple, and its value is in what it forbids. For each study we (1) freeze the model family, training budget, and calendar split before any test date is read. (2) construct features and normalization on the training portion only. (3) enumerate the information views to compare, changing nothing else between them. (4) fit each view under the shared budget, preserving convergence warnings; (5) evaluate every view on the identical held-out events, paired at the event level. (6) score against a fixed metric and, for decisions, a fixed policy and explicit cost model. And (7) apply the pre-registered gate without any outcome-driven adjustment. A failure to certify an effect is recorded as such and never reported as evidence of the opposite effect. This is the whole instrument, and the negatives in this paper are what it returns when it is not allowed to shop for an outcome.

F Data provenance, certification, and the one open endpoint

Sources. All three studies use real Hyperliquid perpetual-futures observations. The forecast-to-decision study draws on native on-change best-bid/offer contracts. The information-sufficiency factorial draws on sampled-L2 receipt-clock snapshots. The reconstruction diagnostic draws on certified BTC breakout and rebound cohorts. Every experiment is gated by a source-and-schema audit that certifies event ordering and receipt timing before the experiment is admitted, so a comparison runs only on inputs whose validity has been checked. Raw market dumps are never redistributed. Each released package ships compact metrics, checksums, and acquisition instructions instead, and no committed path assumes access to the original development machine.

Restricted-data policy. Raw and event-level derivatives are held under the data provider's internal-analysis terms and are excluded from the release. Each package that depends on such data states so and ships a license-clean smoke fixture so that its entry point runs for anyone. Provider receipt time is uncalibrated against exchange time, and the sampled-L2 feed does not establish native queue priority. These are recorded as limitations next to every claim that could be affected.

The open native-FIFO endpoint. The strictest reconstruction test, native-FIFO count tightening on a fully admitted order-lifecycle cohort, is not yet computed, and we state this rather than approximate it. We located the exact public source: an open December 2025 Level-4 order-book archive of the Hyperliquid exchange totaling 195.3 GB, covering BTC, ETH, and SOL, with 744 hourly files per order archive, distributed with raw book diffs and trades. A bounded washout audit over 11 complete hours covered 32,633,820 events and observed 11,328 legacy order identifiers, of which 148 were still first seen in hour 10. Observed-lifecycle coverage reached 99.92985%, but true live-book completeness at the start of the window is unidentified. Because a valid native-FIFO test requires a certified initial book state, we do not claim the endpoint. The decision-relevance conclusion does not depend on it: it rests on the reconstruction diagnostic (Appendix B), which is computed on admitted data.

Provenance records. Table 13 lists the state result identifiers, the independent review identifiers, and the source commit for the six load-bearing packages. Each result was accepted through an independent review before entering the programme, and the reviews are what let us label a package as a valid negative, an inconclusive run, or a blocked endpoint rather than asserting the status ourselves. Every package's full record, task identifiers, protocol revisions, archived implementation branches, and artifact hashes, ships in its `provenance.json`.

G The released experiment programme

The three studies in the main text are the real-data endpoints of a larger programme that we release in full. Table 14 catalogues the thirty curated packages. Each package carries its question, method, compact machine-readable results, selected figures, a provenance record with source identifiers and hashes, and an explicit scientific-status label. The labels are deliberately not collapsed: a completed software check (for example the source-and-schema audit), a valid scientific negative (the forecast-to-decision and information-sufficiency studies), an inconclusive run (the book-washout audit), an invalid cohort, cancelled work, and a blocked empirical objective (the native-FIFO endpoint) are all distinct states, and rounding any of them up to "done" would misrepresent the evidence.

Table 13: Provenance for the six load-bearing packages: state result and independent-review identifiers. Full records, including source commits and artifact hashes, ship in each package's `provenance.json`.

package	results	reviews
Finite-queue ambiguity	R-006, R-013, R-015, R-016	RV-004, RV-009, RV-012, RV-014
Forecast vs. decision	R-007, R-011, R-018, R-037	RV-004, RV-007, RV-013, RV-016
Depth/history/delay	R-008, R-010, R-017, R-019	RV-004, RV-006, RV-013, RV-015
Reconstruction diagnosis	R-034	RV-032
Fixed-model persistence	R-042	RV-040
BTC book-washout audit	R-044	RV-042

Table 14: Released experiment programme: 30 curated packages with execution status. Full narratives, results, provenance, and scientific-status labels ship in the repository. Source: `metadata/experiment-inventory.json`.

IDs	package	execution status
E001	Bounded BTC source, schema, timing, and replay audit	completed
E016,E266	Finite-queue ambiguity on Hyperliquid	completed; blocked on source
E017,E267	Forecast versus decision utility on Hyperliquid	completed
E018,E268	Depth, history, and delay factorial on Hyperliquid	completed
E002	Existing-hour clock-anchor feasibility diagnostic	completed
E200	Causal event-clock and availability certificate	running
E201	Initial-state and matching-rule certificate	running
E202	Chronological day blocks and data-budget feasibility	running
E203	Paired causal coarse/rich observation contract	running
E204	Observed-order labels and execution feasibility	running
E205	Simulator selection and mechanism validation	completed
E210	Controlled action-relevant ambiguity simulation	completed
E253	Breakout/rebound protocol and feasibility	completed
E254	Shared computation and matched-method identity	completed
E255	Label-blind breakout/rebound extraction audit	completed
E256	Restricted-simulator repair and feed validation	completed
E257	Fixed labels and seven-date BTC development cohort	completed
E258	Point versus distributional fill prediction	completed
E259	BTC coarse/predicted-rich/observed-rich comparison	completed
E260	Rich-label novelty and identifiability	completed
E261	Fixed-model BTC August-November evaluation	completed
E262	Cancelled participant checkpoint	cancelled
E263	Contribution triage	completed
E264	Fixed-output reconstruction diagnosis	completed
E265	Recover and admit real Hyperliquid inputs	source candidate found
E269	Prospective fixed-model persistence preparation	completed
E270	Declared BTC persistence source	completed
E271	Fixed-model persistence on one admitted BTC day	completed
E272	Information-value estimand and discriminating next test	completed
E273	BTC unknown-initial-book washout	reviewed terminal prefix

Controlled-mechanism packages. Beyond the three real-data studies, several packages isolate the mechanism behind the negatives in settings where ground truth is known. The controlled action-ambiguity study shows restricted summaries incur 0.075 regret while matched direct and posterior coordinates are equal. The controlled fill-prediction study shows hidden-queue uncertainty matters but reconstruction has no isolated advantage, with direct and distributional outputs coinciding at 512 draws. The fixed-model persistence check on one admitted BTC day contradicts a strict joint persistence condition: the fixed-class contrast is -0.01337 and a uniform predictor beats every learned breakout arm on observed-event loss. These packages are why we present the main-text negatives as mechanisms rather than as isolated empirical accidents.

Status honesty as a feature. We treat the status labels as part of the contribution. A reader who wants only the accepted negatives can filter to the valid packages. A reader who wants to continue the native-FIFO endpoint has the exact source, checksums, and the washout audit that documents what remains unidentified. The programme is designed so that the boundary between what is established and what is open is legible without reading the code.

H CONTROLLED-MECHANISM RESULTS

The controlled studies isolate the mechanism behind the real-data negatives in settings where the hidden state and the optimal action are known, so a null cannot be blamed on market noise.

Action-relevant ambiguity (E-210). Table 15 varies the training size and the label family. Restricted summaries incur a decision regret (for example 0.075 at the smallest universal setting) that shrinks with data, while the matched direct and posterior-restoration coordinates are equal, so no restoration-specific advantage appears. The latent-mean reconstruction MSE moves independently of the decision cost: recovering the latent mean better does not lower the decision cost, which is the controlled analogue of the reconstruction result on market data.

Table 15: Controlled action-ambiguity (E-210): ranked probability score, primary decision cost, decision regret, and latent-mean reconstruction MSE by training size and family. Restricted summaries incur regret while matched coordinates are equal. Source: `experiments/results/e210_finite.csv`.

n_{train}	family	RPS	primary cost	regret	latent MSE
32	universal	0.201384	0.435500	0.010500	1.298229
32	symmetric	0.194937	0.425750	0.000750	1.250000
128	universal	0.196002	0.425656	0.000656	1.267059
128	symmetric	0.193729	0.425000	0.000000	1.250000
512	universal	0.194103	0.425000	0.000000	1.255075
512	symmetric	0.193440	0.425000	0.000000	1.250000

Point versus distributional fill prediction (E-258). Table 16 compares a direct predictor, a point-reconstruction predictor, and a distributional-reconstruction predictor across conditions and training sizes. Hidden-queue uncertainty matters, the point-reconstruction predictor is clearly worse, but the direct and distributional predictors coincide at 512 draws and the direct-compiled predictor matches the distributional one exactly, so reconstruction carries no isolated advantage over predicting the decision directly.

Table 16: Controlled fill-prediction (E-258), selection role: Brier, log loss, calibration error, and rank MSE by condition, training size, and predictor. Direct and distributional predictors coincide at 512 draws; reconstruction shows no isolated advantage. Source: `experiments/results/e258_metrics.csv`.

condition	size	model	Brier	log loss	ECE	rank MSE
informative	128	unconditional_y	0.236403	0.665809	0.063582	–
informative	128	direct_y	0.218317	0.627603	0.063981	–
informative	128	point_y	0.272500	0.756857	0.205469	0.562500
informative	128	distribution_y	0.218936	0.628852	0.065705	0.557968
informative	128	direct_compiled_y	0.218936	0.628852	0.065705	–
informative	128	current_rich_y	0.145427	0.468343	0.135528	–
informative	128	distribution_h	0.213186	0.613884	0.060247	0.548696
informative	128	point_mean_h	0.272500	0.756857	0.205469	0.562500
informative	128	direct_soft_h	0.213186	0.613884	0.060247	–
informative	128	known_mean_point	0.272500	0.756857	0.205469	0.562500
informative	128	exact_coarse_current	0.216287	0.620818	0.073906	–

condition	size	model	Brier	log loss	ECE	rank MSE
informative	128	exact_coarse_source	0.216287	0.620818	0.073906	–
informative	128	exact_rich	0.128750	0.420450	0.012500	–
informative	128	context_only	0.233359	0.658356	0.085156	–
informative	128	exact_unconditional	0.239219	0.671521	0.082813	–
informative	512	unconditional_y	0.237339	0.667713	0.070556	–
informative	512	direct_y	0.210430	0.608710	0.054102	–
informative	512	point_y	0.272500	0.756857	0.205469	0.562500
informative	512	distribution_y	0.210430	0.608710	0.054102	0.546351
informative	512	direct_compiled_y	0.210430	0.608710	0.054102	–
informative	512	current_rich_y	0.129540	0.422807	0.041310	–
informative	512	distribution_h	0.214822	0.617609	0.076785	0.550128
informative	512	point_mean_h	0.272500	0.756857	0.205469	0.562500
informative	512	direct_soft_h	0.214822	0.617609	0.076785	–
informative	512	known_mean_point	0.272500	0.756857	0.205469	0.562500
informative	512	exact_coarse_current	0.216287	0.620818	0.073906	–
informative	512	exact_coarse_source	0.216287	0.620818	0.073906	–
informative	512	exact_rich	0.128750	0.420450	0.012500	–
informative	512	context_only	0.233359	0.658356	0.085156	–
informative	512	exact_unconditional	0.239219	0.671521	0.082813	–
zero	128	unconditional_y	0.260971	0.716023	0.107332	–
zero	128	direct_y	0.233780	0.660119	0.099132	–
zero	128	point_y	0.227187	0.658831	0.111719	0.460938
zero	128	distribution_y	0.233780	0.660119	0.099132	0.540339
zero	128	direct_compiled_y	0.233780	0.660119	0.099132	–
zero	128	current_rich_y	0.141311	0.449890	0.092970	–
zero	128	distribution_h	0.219565	0.630344	0.074597	0.485886
zero	128	point_mean_h	0.227187	0.658831	0.111719	0.460938
zero	128	direct_soft_h	0.219565	0.630344	0.074597	–
zero	128	known_mean_point	0.227187	0.658831	0.111719	0.460938
zero	128	exact_coarse_current	0.216172	0.623795	0.038281	–
zero	128	exact_coarse_source	0.216172	0.623795	0.038281	–
zero	128	exact_rich	0.125098	0.411581	0.034766	–
zero	128	context_only	0.216172	0.623795	0.038281	–
zero	128	exact_unconditional	0.250156	0.693469	0.026562	–
zero	512	unconditional_y	0.251807	0.696824	0.048547	–
zero	512	direct_y	0.219139	0.629516	0.062349	–
zero	512	point_y	0.227187	0.658831	0.111719	0.460938
zero	512	distribution_y	0.219139	0.629516	0.062349	0.475028
zero	512	direct_compiled_y	0.219139	0.629516	0.062349	–
zero	512	current_rich_y	0.130753	0.428611	0.098630	–
zero	512	distribution_h	0.216958	0.625334	0.041112	0.466307
zero	512	point_mean_h	0.227187	0.658831	0.111719	0.460938
zero	512	direct_soft_h	0.216958	0.625334	0.041112	–
zero	512	known_mean_point	0.227187	0.658831	0.111719	0.460938
zero	512	exact_coarse_current	0.216172	0.623795	0.038281	–
zero	512	exact_coarse_source	0.216172	0.623795	0.038281	–
zero	512	exact_rich	0.125098	0.411581	0.034766	–
zero	512	context_only	0.216172	0.623795	0.038281	–
zero	512	exact_unconditional	0.250156	0.693469	0.026562	–

Fixed-model persistence (E-271). Table 17 reports the persistence check on one admitted BTC day. The strict joint condition is contradicted: raw and favorable-class contrasts are positive while the fixed-class contrast is -0.01337, and a uniform predictor beats every learned breakout arm on observed-event loss. One source-exposed day cannot support a population claim, but it is consistent with the reconstruction and sufficiency negatives: the richer, learned arm does not persist as the better decision.

Table 17: Fixed-model persistence (E-271) on BTC 2026-02-01: per-endpoint values by family, kind, and metric. The fixed-class contrast is negative while raw and favorable-class contrasts are positive. Source: `experiments/results/e271_main.csv`.

family	kind	metric	endpoint	value	status	n
breakout	method	C	raw	0.377905	evaluable	880
breakout	method	C	frozen_mixture	0.303310	evaluable	880
breakout	method	C	F	0.277484	evaluable	880
breakout	method	C	A	0.618632	evaluable	880
breakout	method	C	N	0.243273	evaluable	880
breakout	method	P	raw	0.376674	evaluable	880
breakout	method	P	frozen_mixture	0.295825	evaluable	880
breakout	method	P	F	0.293166	evaluable	880
breakout	method	P	A	0.610970	evaluable	880
breakout	method	P	N	0.220948	evaluable	880
breakout	method	R	raw	0.378487	evaluable	880
breakout	method	R	frozen_mixture	0.289943	evaluable	880
breakout	method	R	F	0.294158	evaluable	880
breakout	method	R	A	0.627595	evaluable	880
breakout	method	R	N	0.205180	evaluable	880
breakout	method	onehot-F	raw	0.568182	evaluable	880
breakout	method	onehot-F	frozen_mixture	0.659259	evaluable	880
breakout	method	onehot-F	F	0.000000	evaluable	880
breakout	method	onehot-F	A	1.000000	evaluable	880
breakout	method	onehot-F	N	1.000000	evaluable	880
breakout	method	onehot-A	raw	0.680682	evaluable	880
breakout	method	onehot-A	frozen_mixture	0.871111	evaluable	880
breakout	method	onehot-A	F	1.000000	evaluable	880
breakout	method	onehot-A	A	0.000000	evaluable	880
breakout	method	onehot-A	N	1.000000	evaluable	880
breakout	method	onehot-N	raw	0.751136	evaluable	880
breakout	method	onehot-N	frozen_mixture	0.469630	evaluable	880
breakout	method	onehot-N	F	1.000000	evaluable	880
breakout	method	onehot-N	A	1.000000	evaluable	880
breakout	method	onehot-N	N	0.000000	evaluable	880
breakout	method	uniform	raw	0.333333	evaluable	880
breakout	method	uniform	frozen_mixture	0.333333	evaluable	880
breakout	method	uniform	F	0.333333	evaluable	880
breakout	method	uniform	A	0.333333	evaluable	880
breakout	method	uniform	N	0.333333	evaluable	880
breakout	method	train-frequency	raw	0.386720	evaluable	880
breakout	method	train-frequency	frozen_mixture	0.292995	evaluable	880
breakout	method	train-frequency	F	0.366264	evaluable	880
breakout	method	train-frequency	A	0.578116	evaluable	880
breakout	method	train-frequency	N	0.176634	evaluable	880
breakout	contrast	P-C	raw	-0.001231	evaluable	880
breakout	contrast	P-C	frozen_mixture	-0.007485	evaluable	880
breakout	contrast	P-C	F	0.015682	evaluable	880
breakout	contrast	P-C	A	-0.007662	evaluable	880
breakout	contrast	P-C	N	-0.022325	evaluable	880
breakout	contrast	R-C	raw	0.000582	evaluable	880
breakout	contrast	R-C	frozen_mixture	-0.013367	evaluable	880
breakout	contrast	R-C	F	0.016674	evaluable	880
breakout	contrast	R-C	A	0.008963	evaluable	880
breakout	contrast	R-C	N	-0.038093	evaluable	880
breakout	contrast	R-P	raw	0.001813	evaluable	880
breakout	contrast	R-P	frozen_mixture	-0.005882	evaluable	880
breakout	contrast	R-P	F	0.000992	evaluable	880
breakout	contrast	R-P	A	0.016625	evaluable	880
breakout	contrast	R-P	N	-0.015768	evaluable	880
breakout	contrast	C-frequency	raw	-0.008815	evaluable	880
breakout	contrast	C-frequency	frozen_mixture	0.010314	evaluable	880
breakout	contrast	C-frequency	F	-0.088780	evaluable	880

family	kind	metric	endpoint	value	status	n
breakout	contrast	C-frequency	A	0.040516	evaluable	880
breakout	contrast	C-frequency	N	0.066639	evaluable	880
breakout	contrast	P-frequency	raw	-0.010046	evaluable	880
breakout	contrast	P-frequency	frozen_mixture	0.002830	evaluable	880
breakout	contrast	P-frequency	F	-0.073098	evaluable	880
breakout	contrast	P-frequency	A	0.032854	evaluable	880
breakout	contrast	P-frequency	N	0.044314	evaluable	880
breakout	contrast	R-frequency	raw	-0.008233	evaluable	880
breakout	contrast	R-frequency	frozen_mixture	-0.003053	evaluable	880
breakout	contrast	R-frequency	F	-0.072106	evaluable	880
breakout	contrast	R-frequency	A	0.049479	evaluable	880
breakout	contrast	R-frequency	N	0.028546	evaluable	880
breakout	contrast	frequency-uniform	raw	0.053387	evaluable	880
breakout	contrast	frequency-uniform	frozen_mixture	-0.040338	evaluable	880
breakout	contrast	frequency-uniform	F	0.032931	evaluable	880
breakout	contrast	frequency-uniform	A	0.244782	evaluable	880
breakout	contrast	frequency-uniform	N	-0.156699	evaluable	880
rebound	method	C	raw	0.339716	evaluable	865
rebound	method	C	frozen_mixture	0.310577	evaluable	865
rebound	method	C	F	0.711610	evaluable	865
rebound	method	C	A	0.197967	evaluable	865
rebound	method	C	N	0.359078	evaluable	865
rebound	method	P	raw	0.341107	evaluable	865
rebound	method	P	frozen_mixture	0.310627	evaluable	865
rebound	method	P	F	0.702873	evaluable	865
rebound	method	P	A	0.207901	evaluable	865
rebound	method	P	N	0.350677	evaluable	865
rebound	method	R	raw	0.366120	evaluable	865
rebound	method	R	frozen_mixture	0.318122	evaluable	865
rebound	method	R	F	0.773648	evaluable	865
rebound	method	R	A	0.246041	evaluable	865
rebound	method	R	N	0.317585	evaluable	865
rebound	method	onehot-F	raw	0.809249	evaluable	865
rebound	method	onehot-F	frozen_mixture	0.926298	evaluable	865
rebound	method	onehot-F	F	0.000000	evaluable	865
rebound	method	onehot-F	A	1.000000	evaluable	865
rebound	method	onehot-F	N	1.000000	evaluable	865
rebound	method	onehot-A	raw	0.462428	evaluable	865
rebound	method	onehot-A	frozen_mixture	0.537688	evaluable	865
rebound	method	onehot-A	F	1.000000	evaluable	865
rebound	method	onehot-A	A	0.000000	evaluable	865
rebound	method	onehot-A	N	1.000000	evaluable	865
rebound	method	onehot-N	raw	0.728324	evaluable	865
rebound	method	onehot-N	frozen_mixture	0.536013	evaluable	865
rebound	method	onehot-N	F	1.000000	evaluable	865
rebound	method	onehot-N	A	1.000000	evaluable	865
rebound	method	onehot-N	N	0.000000	evaluable	865
rebound	method	uniform	raw	0.333333	evaluable	865
rebound	method	uniform	frozen_mixture	0.333333	evaluable	865
rebound	method	uniform	F	0.333333	evaluable	865
rebound	method	uniform	A	0.333333	evaluable	865
rebound	method	uniform	N	0.333333	evaluable	865
rebound	method	train-frequency	raw	0.328585	evaluable	865
rebound	method	train-frequency	frozen_mixture	0.282776	evaluable	865
rebound	method	train-frequency	F	0.643522	evaluable	865
rebound	method	train-frequency	A	0.254912	evaluable	865
rebound	method	train-frequency	N	0.253237	evaluable	865
rebound	contrast	P-C	raw	0.001391	evaluable	865
rebound	contrast	P-C	frozen_mixture	0.000050	evaluable	865
rebound	contrast	P-C	F	-0.008737	evaluable	865
rebound	contrast	P-C	A	0.009934	evaluable	865
rebound	contrast	P-C	N	-0.008401	evaluable	865
rebound	contrast	R-C	raw	0.026404	evaluable	865

family	kind	metric	endpoint	value	status	n
rebound	contrast	R-C	frozen_mixture	0.007545	evaluable	865
rebound	contrast	R-C	F	0.062038	evaluable	865
rebound	contrast	R-C	A	0.048073	evaluable	865
rebound	contrast	R-C	N	-0.041493	evaluable	865
rebound	contrast	R-P	raw	0.025013	evaluable	865
rebound	contrast	R-P	frozen_mixture	0.007495	evaluable	865
rebound	contrast	R-P	F	0.070775	evaluable	865
rebound	contrast	R-P	A	0.038140	evaluable	865
rebound	contrast	R-P	N	-0.033092	evaluable	865
rebound	contrast	C-frequency	raw	0.011131	evaluable	865
rebound	contrast	C-frequency	frozen_mixture	0.027801	evaluable	865
rebound	contrast	C-frequency	F	0.068088	evaluable	865
rebound	contrast	C-frequency	A	-0.056945	evaluable	865
rebound	contrast	C-frequency	N	0.105841	evaluable	865
rebound	contrast	P-frequency	raw	0.012522	evaluable	865
rebound	contrast	P-frequency	frozen_mixture	0.027851	evaluable	865
rebound	contrast	P-frequency	F	0.059351	evaluable	865
rebound	contrast	P-frequency	A	-0.047011	evaluable	865
rebound	contrast	P-frequency	N	0.097440	evaluable	865
rebound	contrast	R-frequency	raw	0.037535	evaluable	865
rebound	contrast	R-frequency	frozen_mixture	0.035346	evaluable	865
rebound	contrast	R-frequency	F	0.130126	evaluable	865
rebound	contrast	R-frequency	A	-0.008871	evaluable	865
rebound	contrast	R-frequency	N	0.064348	evaluable	865
rebound	contrast	frequency-uniform	raw	-0.004748	evaluable	865
rebound	contrast	frequency-uniform	frozen_mixture	-0.050557	evaluable	865
rebound	contrast	frequency-uniform	F	0.310189	evaluable	865
rebound	contrast	frequency-uniform	A	-0.078421	evaluable	865
rebound	contrast	frequency-uniform	N	-0.080096	evaluable	865
breakout	joint_component	R-C	raw	0.000582	evaluable	880
breakout	joint_component	R-C	fixed_class	-0.013367	evaluable	880
breakout	joint_component	R-C	favorable_conditional	0.016674	evaluable	880

I REPRODUCIBILITY

Environment. All experiments run on a laptop-class machine: macOS, CPU or Apple MPS only, no CUDA, no large-scale training. Models are small by design (linear, gradient boosting, and a compact multilayer perceptron), and the full suite fits within a modest time budget. Randomness is controlled with fixed seeds. Where a noise axis is declared (training seeds, item bootstrap, or a fixed grid), the corresponding uncertainty is reported rather than a single point.

Entry points. Each package exposes the shortest truthful entry point (`run.py`) and a frozen configuration. A license-clean smoke fixture lets every package run end to end without the restricted raw data, so the code path is verifiable by anyone. The full result requires the acquisition steps recorded in each package's provenance file. Numbers that appear in this paper are regenerated from the staged result files under `experiments/results/`, and every table in the appendix is emitted directly from those files so that a printed value always equals a source value. An anonymized copy of the full programme, its results, and this paper is available for review at https://anonymous.4open.science/r/learning-with-less-C713.

Checklist.

- **Data.** Sources named with identifiers and checksums. Raw dumps excluded. Restricted terms disclosed. Smoke fixtures license-clean.
- **Splits.** Train/validation/test dates fixed before any test date is scored. Feature construction and normalization training-only.
- **Models.** Families and budgets shared across information views. Neural iteration caps recorded, warnings preserved.

- **Evaluation.** Metrics and decision policies pre-registered. Joint gates frozen. No outcome-driven tweaks. Failure-to-certify never reported as proof of the opposite.
- **Artifacts.** Compact results, figures, provenance, and scientific-status labels released per package. Figures regenerate from result files only.

Scope statement. The released evidence supports the scoped negatives and the controlled mechanisms exactly as labeled. It does not support claims of live-trading profitability, calibrated latency, native queue priority, or generalization beyond the venues, horizons, and assets tested. The one open endpoint (native-FIFO count tightening on a fully admitted cohort) is flagged with its exact source so that it can be completed rather than assumed.

J DATASHEET FOR THE RELEASED PROGRAMME

We summarize the released artifacts in the datasheet style, so a reader can judge fitness for reuse before touching the code.

Motivation. The programme exists to test, as controlled interventions on real market data, whether more granular or more accurately reconstructed order-book information improves decisions. It was built to keep scientific status honest: negatives, controlled mechanisms, inconclusive runs, and blocked endpoints are all represented rather than filtered to positive findings.

Composition. The release contains thirty curated experiment packages (Table 14) covering three real-data studies (hidden-structure reconstruction, forecast-to-decision value, and information sufficiency), controlled simulator mechanisms, breakout/rebound cohorts, temporal-persistence checks, and source-and-schema certification. Each package holds a question statement, a method description, compact machine-readable results, selected figures, a provenance record, and a scientific-status label. Numeric results in this paper are staged under `experiments/results/` and every table is generated directly from those files.

Collection process. Market observations were acquired from Hyperliquid perpetual-futures feeds under the provider's terms. Native on-change best-bid/offer data back the forecast-to-decision study. Sampled-L2 receipt-clock snapshots back the information-sufficiency factorial. Certified BTC breakout and rebound cohorts back the reconstruction diagnostic. Every experiment is admitted only after a source-and-schema audit certifies event ordering and receipt timing.

Preprocessing and labeling. Features and normalization are fit on training portions only and frozen before any test date is scored. Targets are past-only definitions of short-horizon moves or fill/adverse/no-move outcomes. No test-date information enters feature construction, and no per-view retuning is permitted.

Distribution and licensing. Code is released under a permissive license. Raw market dumps and event-level derivatives are excluded under the provider's internal-analysis terms and represented by checksums and acquisition instructions. Each package ships a license-clean smoke fixture so its entry point runs without the restricted data. No committed path assumes access to the original development machine, and secrets remain in ignored local configuration.

Maintenance and known gaps. The one blocked endpoint, native-FIFO count tightening on a fully admitted order-lifecycle cohort, is documented with its exact public source (a 195.3 GB December 2025 Level-4 Hyperliquid archive) and a bounded washout audit, so it can be completed by a future user. It is labeled blocked, not done.

K THREATS TO VALIDITY

We enumerate the threats to each conclusion and the design choice that addresses it, so a reviewer can check that the negatives are not artifacts.

Outcome shopping. The central threat to any negative is that the analyst kept searching until the result appeared. We freeze every plan, model families, budgets, splits, policies, and gates, before test dates are read, and we forbid outcome-driven tweaks. A failure to certify an effect is recorded as such and never reported as proof of the opposite. The soundness of a negative depends on this, and it is enforced by construction rather than by assertion.

Model capacity confound. A view could look uninformative only because the model was too weak to use it. We hold the model family and budget constant across views and compare views within each family, and in the information-sufficiency study we add an optimizer diagnostic that varies the neural budget. The asset-dependent direction survives the budget change, so the sufficiency negative is not a pure underfitting artifact. Neural cells that hit their iteration cap are flagged and never used as the sole basis for a claim.

Metric and policy dependence. A decision negative could reflect one unlucky metric or policy. The forecast-to-decision study uses three policies and sweeps cost and latency scenarios. The argmax negative is absolute (even the oracle loses), the confidence policy abstains, and the scheduling policy straddles zero across scenarios. The conclusion is stable across the policy set, not a single strategy.

Aggregation artifacts. An aggregate over heterogeneous cohorts can manufacture or hide an effect. We report gaps under a fixed weighting against explicit baselines and decompose each gap into mixture and within-class terms (Appendix E). The mixture term carries the earlier apparent depth advantage, which is why we do not report a single aggregate.

Data realism. Uncalibrated receipt time, sampled-L2 depth, and quote-based utility limit external validity: the forecast-to-decision result is a decision-utility statement, not realized profitability, and the information-sufficiency study's small date panel rejects a simple rule rather than proving depth never helps. We state these limits next to each claim and withhold any claim whose inputs are not certified, most visibly the native-FIFO endpoint.

Generalization. The negatives are scoped to the venues, horizons, and assets tested. We do not claim they transfer to other markets, and the released protocol is exactly what a reader would rerun to test transfer elsewhere.

L Extended discussion of related literature

This appendix expands the six clusters of Section 2 and states, cluster by cluster, what the prior line establishes and where our matched measurement adds evidence.

Predictive modeling of the book. The dominant line treats the limit order book as an input to a predictor and asks how much signal a model can extract, from convolutional and recurrent architectures (Sirignano, 2016; Zhang et al., 2018c; Nousi et al., 2018; Zhang et al., 2018b) through attention and persistence-aware designs (Jung & Lee, 2024; Briola et al., 2024b; Lucchese et al., 2022; Briola et al., 2024a) to generative and foundation models (Zhang et al., 2026b; 2021; Backhouse et al., 2025) and representation-learning surveys and benchmarks (Zhong et al., 2025; Ntakaris et al., 2017; Nagy et al., 2025; Cao et al., 2022). This line reports predictive scores. Our studies hold the model fixed and report the decision, which is why an improvement that this line would count as progress can be a null or a loss in our measurement.

Order-flow structure, queues, and impact. A second line models the mechanics by which book detail could carry decision value: multi-level and generalized order-flow imbalance (Xu et al., 2019; Cont et al., 2010; Su et al., 2021; Hu & Zhang, 2025; Lipton et al., 2013), queue and order-position dynamics (Guo et al., 2015; Cont & Larrard, 2012; Toke, 2013; Toke & Yoshida, 2016; Cont & Mueller, 2019), resiliency and depth after events (Bechler & Ludkovski, 2017; Xu et al., 2016; Krause et al., 2021), and price impact of orders and flow (Said et al., 2018; Brokmann et al., 2014; Lemhadri, 2018; Huang et al., 2019). These establish that finer structure is informative in a statistical sense. Our reconstruction and information-sufficiency studies test whether that information changes a decision and find that on real data it usually does not.

Fill probability and execution. Work on fill probability and execution supplies the concrete channel through which queues matter: survival and state-dependent fill models (Arroyo et al., 2023; Lokin & Yu, 2024; Guo et al., 2015) and optimal-execution and market-making theory (Kato, 2014; 2009; Curato et al., 2014; Bechler & Ludkovski, 2014; Chavez-Casillas et al., 2024; Baldacci & Bergault, 2021). Our finite-queue analysis confirms the channel exists but is rare, and the reconstruction diagnostic shows recovering the queue does not move the fill/adverse/no-move decision.

Forecast value versus decision value. The line closest to our thesis documents that accuracy and decision quality diverge: rank-correlation breakdowns (Falezza, 2026; Raeth & Ludwig, 2025), economic-value gaps beyond error metrics (Maciejowska et al., 2025; Kath & Ziel, 2018; Hirsch & Ziel, 2026; Walid et al., 2025), decision-focused learning (Wilder et al., 2018; Zhou et al., 2024; Silvestri et al., 2023; Islam & Wasi, 2023; Tankov & Tinsi, 2021), and information-driven trading value (Samuel, 2020). Proper guarantees exist under matched rules and liquidity (Gu et al., 2026; Dubach, 2026). The forecast-to-decision study supplies the market measurement with a frozen joint gate.

Forecast evaluation and calibration. Tools that evaluate forecasts across decision thresholds (Ziegel et al., 2017; Dimitriadis et al., 2023; Ehm et al., 2015; Taggart, 2021; Allen et al., 2024; Bouallegue et al., 2015) and calibration and communication of probabilistic forecasts (Graziani et al., 2019; Loveday et al., 2023; Raftery, 2014; Crisostomo & Couso, 2018) already argue that a single score is a weak decision surrogate. We quantify the resulting disagreement in basis points under an explicit policy and cost model.

Benchmarks, leakage, sufficiency, and simulators. Recent benchmark and audit work targets decision-time leakage and execution realism (Zhang et al., 2026a; Yao & Zheng, 2026; Cheng et al., 2026; Wang et al., 2025), backtest correctness and overfitting (Low et al., 2015; Koshiyama & Firoozye, 2019; Pham et al., 2026; Sun & Lyuu, 2022; Maier-Paape & Platen, 2014), and agent-based simulators for counterfactual evaluation (Dwarakanath et al., 2024; Vyetrenko & Xu, 2019). On the theory side, decision-sufficient representations and information-theoretic limits bound what any observable can buy (Ye et al., 2026; Walsh, 2026; Catt, 2026; Yoon & Kim, 2026). Our released programme operationalizes this stance with matched interventions, frozen gates, source certification, and honest status labels.

M LITERATURE INVENTORY AND COVERAGE

The programme is positioned against a bibliography of verified references spanning limit-order-book forecasting, order-flow and queue modeling, execution and market impact, forecast-versus-decision value, forecast evaluation and calibration, and benchmark, leakage, and sufficiency work. Table 18 lists every entry with its role and its citation count at retrieval time, so the inventory doubles as a coverage record: each cited work maps to at least one cluster in Section 2. We include the full inventory because a synthesis of negatives is only credible if it is read against the strongest prior claims of positive information value, and the inventory makes that reading auditable.

Table 18: Literature inventory: every verified reference in the programme's bibliography, grouped by role, with citation counts at retrieval time. Each entry is cited here so the inventory doubles as a coverage record.

role	year	work	cites
foundational	2018	Melding the Data-Decisions Pipeline: Decision-Focused Learning for Com (Wilder et al., 2018)	122
foundational	2017	Benchmark Dataset for Mid-Price Forecasting of Limit Order Book Data w (Ntakaris et al., 2017)	116
foundational	2016	Deep Learning for Limit Order Books (Sirignano, 2016)	103
foundational	2018	The value of forecasts: Quantifying the economic gains of accurate qua (Kath & Ziel, 2018)	68
foundational	2014	Optimal execution with nonlinear transient market impact (Curato et al., 2014)	46

role	year	work	cites
foundational	2014	Optimal Execution with Dynamic Order Flow Imbalance (Bechler & Ludkovski, 2014)	32
foundational	2014	VWAP Execution as an Optimal Strategy (Kato, 2014)	15
foundational	2009	An Optimal Execution Problem with Market Impact (Kato, 2009)	14
foundational	2007	Correlation of coming limit price with order book in stock markets (ichi Maskawa, 2007)	11
foundational	2018	Market Impact: A Systematic Study of Limit Orders (Said et al., 2018)	8
foundational	2015	Dynamics of Order Positions and Related Queues in a Limit Order Book (Guo et al., 2015)	6
foundational	2021	Forecasting Quoted Depth With the Limit Order Book (Libman et al., 2021)	6
foundational	2025	Statistical and economic evaluation of forecasts in electricity market (Maciejowska et al., 2025)	3
foundational	2004	Uncertainty relations in models of market microstructure (Theodosopoulos, 2004)	3
foundational	2026	When Forecast Accuracy Fails: Rank Correlation and Decision Quality in (Falezza, 2026)	0
foundational	2026	Probabilistic Forecasting for Day-ahead Electricity Prices, Battery Tr (Hirsch & Ziel, 2026)	0
foundational	2025	Forecast Skill Is Not Decision Skill: Evidence from Weather-Dependent (Raeth & Ludwig, 2025)	0
foundational	2018	Machine Learning for Forecasting Mid Price Movement using Limit Order (Nousi et al., 2018)	0
foundational	2015	Correctness of Backtest Engines (Low et al., 2015)	0
foundational	2014	Backtest of Trading Systems on Candle Charts (Maier-Paape & Platen, 2014)	0
foundational	2026	The Anatomy of a Decentralized Prediction Market: Microstructure Evide (Dubach, 2026)	0
foundational	2018	BDLOB: Bayesian Deep Convolutional Neural Networks for Limit Order Boo (Zhang et al., 2018b)	0
sota	2010	The Price Impact of Order Book Events (Cont et al., 2010)	297
sota	2018	DeepLOB: Deep Convolutional Neural Networks for Limit Order Books (Zhang et al., 2018c)	238
sota	2015	Of Quantiles and Expectiles: Consistent Scoring Functions, Choquet Rep (Ehm et al., 2015)	151
sota	2021	Deep Learning for Market by Order Data (Zhang et al., 2021)	30
sota	2023	Comparative Analysis of Machine Learning, Hybrid, and Deep Learning Fo (Ampountolas, 2023)	28
sota	2023	Deep Attentive Survival Analysis in Limit Order Books: Estimating Fill (Arroyo et al., 2023)	18
sota	2012	Price Jump Prediction in Limit Order Book (Zheng et al., 2012)	15
sota	2023	Evaluating Probabilistic Classifiers: The Triptych (Dimitriadis et al., 2023)	14
sota	2022	The Short-Term Predictability of Returns in Order Book Markets: a Deep (Lucchese et al., 2022)	14
sota	2019	Multi-Level Order-Flow Imbalance in a Limit Order Book (Xu et al., 2019)	11
sota	2024	HLOB – Information Persistence and Structure in Limit Order Books (Briola et al., 2024b)	9
sota	2019	Probabilistic Recalibration of Forecasts (Graziani et al., 2019)	2
sota	2024	Attention-Based Reading, Highlighting, and Forecasting of the Limit Or (Jung & Lee, 2024)	1
sota	2026	When do prophets profit in prediction markets? (Gu et al., 2026)	0
sota	2026	Beyond Agent Architecture: Execution Assumptions and Reproducibility i (Yao & Zheng, 2026)	0
sota	2026	When Alpha Disappears: A One-Switch Benchmark for Decision-Time Leakag (Zhang et al., 2026a)	0
sota	2026	PolyBench: Benchmarking LLM Forecasting and Trading Capabilities on Li (Cheng et al., 2026)	0
sota	2017	Murphy Diagrams: Forecast Evaluation of Expected Shortfall (Ziegel et al., 2017)	0
sota	2019	Machine Learning on EPEX Order Books: Insights and Forecasts (Schnurch & Wagner, 2019)	0

role	year	work	cites
sota	2022	Deep Reinforcement Learning for Cryptocurrency Trading: Practical Appr (Gort et al., 2022)	0
sota	2024	Deep Limit Order Book Forecasting (Briola et al., 2024a)	0
sota	2025	Economic and Reliability Value of Improved Offshore Wind Forecasting i (Walid et al., 2025)	0
sota	2025	LOB-Bench: Benchmarking Generative AI for Finance – an Application to (Nagy et al., 2025)	0
sota	2024	Fill Probabilities in a Limit Order Book with State-Dependent Stochast (Lokin & Yu, 2024)	0
sota	2026	In-Network Market Prediction Using Machine Learning and Limit Order Bo (Hong et al., 2026)	0
sota	2026	M3: A State-Event Generative Foundation Model for Market Microstructur (Zhang et al., 2026b)	0
sota	2023	Integrating Tick-level Data and Periodical Signal for High-frequency M (He et al., 2023)	0
sota	2019	Risk-Sensitive Compact Decision Trees for Autonomous Execution in Pres (Vyetrenko & Xu, 2019)	0
sota	2021	The Price Impact of Generalized Order Flow Imbalance (Su et al., 2021)	0
sota	2025	Stochastic Price Dynamics in Response to Order Flow Imbalance: Evidenc (Hu & Zhang, 2025)	0
sota	2024	Adaptive Optimal Market Making Strategies with Inventory Liquidation C (Chavez-Casillas et al., 2024)	0
adjacent	2014	Use and Communication of Probabilistic Forecasts (Raftery, 2014)	52
adjacent	2014	Slow decay of impact in equity markets (Brokmann et al., 2014)	40
adjacent	2023	A probabilistic forecast methodology for volatile electricity prices i (Cornell et al., 2023)	36
adjacent	2012	Order book dynamics in liquid markets: limit theorems and diffusion ap (Cont & Larrard, 2012)	36
adjacent	2021	Point forecasting and forecast evaluation with generalized Huber loss (Taggart, 2021)	27
adjacent	2020	Information Token Driven Machine Learning for Electronic Markets: Perf (Samuel, 2020)	23
adjacent	2013	The order book as a queueing system: average depth and influence of th (Toke, 2013)	22
adjacent	2016	Modelling intensities of order flows in a limit order book (Toke & Yoshida, 2016)	20
adjacent	2010	An algorithmic information-theoretic approach to the behaviour of fina (Zenil & Delahaye, 2010)	20
adjacent	2019	A stochastic partial differential equation model for limit order book (Cont & Mueller, 2019)	19
adjacent	2015	Quantile forecast discrimination ability and value (Bouallegue et al., 2015)	14
adjacent	2023	Optimal Sampling Designs for Multi-dimensional Streaming Time Series w (Xie et al., 2023)	13
adjacent	2014	Hydrodynamic limit of order book dynamics (Gao & Deng, 2014)	12
adjacent	2018	Financial density forecasts: A comprehensive comparison of risk-neutra (Crisostomo & Couso, 2018)	11
adjacent	2020	Beating the market with a bad predictive model (Hubacek & Sir, 2020)	9
adjacent	2022	Bidding and Scheduling in Energy Markets: Which Probabilistic Forecast (Beykirch et al., 2022)	8
adjacent	2020	Probabilistic multivariate electricity price forecasting using implici (Janke & Steinke, 2020)	8
adjacent	2024	Decision Focused Causal Learning for Direct Counterfactual Marketing O (Zhou et al., 2024)	7
adjacent	2015	Quasi-Centralized Limit Order Books (Gould et al., 2015)	7
adjacent	2017	Order Flows and Limit Order Book Resiliency on the Meso-Scale (Bechler & Ludkovski, 2017)	6
adjacent	2015	An Information Theoretic Analysis of Sequential Decision-Making (Dorpinghaus et al., 2015)	6
adjacent	2021	Contextual Sentence Analysis for the Sentiment Prediction on Financial (Pontes & Benjannet, 2021)	6
adjacent	2023	A User-Focused Approach to Evaluating Probabilistic and Categorical Fo (Loveday et al., 2023)	5

role	year	work	cites
adjacent	2024	Large Language Models for Financial Aid in Financial Time-series Forec (Islam et al., 2024)	5
adjacent	2018	An Adaptive Tabu Search Algorithm for Market Clearing Problem in Turki (Kurt et al., 2018)	5
adjacent	2019	From Glosten-Milgrom to the whole limit order book and applications to (Huang et al., 2019)	4
adjacent	2016	Limit-order book resiliency after effective market orders: Spread, dep (Xu et al., 2016)	3
adjacent	2017	Empirical comparison of three models for determining market clearing p (Ceyhan et al., 2017)	3
adjacent	2021	Optimal incentives in a limit order book: a SPDE control approach (Baldacci & Bergault, 2021)	2
adjacent	2023	Score Function Gradient Estimation to Widen the Applicability of Decis (Silvestri et al., 2023)	2
adjacent	2024	Tail calibration of probabilistic forecasts (Allen et al., 2024)	2
adjacent	2021	Two Price Regimes in Limit Order Books: Liquidity Cushion and Fragment (Krause et al., 2021)	1
adjacent	2020	Limit Order Book (LOB) shape modeling in presence of heterogeneously i (Drame, 2020)	1
adjacent	2025	Limit Order Book Dynamics in Matching Markets: Microstructure, Spread, (Wu, 2025)	0
adjacent	2026	Decomposing Crowd Wisdom: Domain-Specific Calibration Dynamics in Pred (Le, 2026)	0
adjacent	2025	FinTSBridge: A New Evaluation Suite for Real-world Financial Predictio (Wang et al., 2025)	0
adjacent	2025	An End-to-End Approach for Microgrid Probabilistic Forecasting and Rob (Cao & Xu, 2025)	0
adjacent	2021	Forecasting high-frequency financial time series: an adaptive learning (Yang, 2021)	0
adjacent	2026	Early Detection of Latent Microstructure Regimes in Limit Order Books (Hiremath & Hiremath, 2026)	0
adjacent	2023	Hawkes-based cryptocurrency forecasting via Limit Order Book data (Cestari et al., 2023)	0
adjacent	2025	Painting the market: generative diffusion models for financial limit o (Backhouse et al., 2025)	0
adjacent	2021	Online Trading Models with Deep Reinforcement Learning in the Forex Ma (Ishikawa & Nakata, 2021)	0
adjacent	2018	Market Impact in a Latent Order Book (Lemhadri, 2018)	0
adjacent	2013	Trade arrival dynamics and quote imbalance in a limit order book (Lipton et al., 2013)	0
adjacent	2025	Representation Learning of Limit Order Book: A Comprehensive Study and (Zhong et al., 2025)	0
adjacent	2026	AlgoXpert Alpha Research Framework. A Rigorous IS WFA OOS Protocol for (Pham et al., 2026)	0
adjacent	2019	Avoiding Backtesting Overfitting by Covariance-Penalties: an empirical (Koshiyama & Firoozye, 2019)	0
adjacent	2022	Backtesting Trading Strategies with GAN To Avoid Overfitting (Sun & Lyuu, 2022)	0
adjacent	2021	Decision making with dynamic probabilistic forecasts (Tankov & Tinsi, 2021)	0
adjacent	2022	DSLOB: A Synthetic Limit Order Book Dataset for Benchmarking Forecasti (Cao et al., 2022)	0
adjacent	2024	ABIDES-Economist: Agent-Based Simulator of Economic Systems with Learn (Dwarakanath et al., 2024)	0
adjacent	2008	Statistical Arbitrage and Optimal Trading with Transaction Costs in Fu (Tsagaris, 2008)	0
adjacent	2023	Optimizing Inventory Routing: A Decision-Focused Learning Approach usi (Islam & Wasi, 2023)	0
adjacent	2022	E-backtesting (Wang et al., 2022)	0
adjacent	2018	Benchmarking Deep Sequential Models on Volatility Predictions for Fina (Zhang et al., 2018a)	0
adjacent	2012	Isoelastic Agents and Wealth Updates in Machine Learning Markets (Storkey et al., 2012)	0

role	year	work	cites
adjacent	2026	Forecastability as an Information-Theoretic Limit on Prediction (Catt, 2026)	0
adjacent	2016	Market Dynamics vs. Statistics: Limit Order Book Example (Malyshkin & Bakhramov, 2016)	0
adjacent	2026	Finite Reliability Representations: Noise-Calibrated Belief-Space Cove (Yoon & Kim, 2026)	0
adjacent	2011	Stochastic Price Dynamics Implied By the Limit Order Book (Langnau & Punchev, 2011)	0
adjacent	2014	One-level limit order book models with memory and variable spread (Chavez-Casillas & Figueroa-Lopez, 2014)	0
adjacent	2020	Informed trading, limit order book and implementation shortfall: equil (Cetin & Waelbroeck, 2020)	0
adjacent	2025	Agent-based model of information diffusion in the limit order book tra (Wilinski & Kanniainen, 2025)	0
adjacent	2026	What Happens When Institutional Liquidity Enters Prediction Markets: I (Dalen, 2026)	0
adjacent	2026	Volatility Forecasting and Return Prediction under Market Regimes: Evi (Fang & Slepaczuk, 2026)	0
adjacent	2009	Pragmatic Information Rates, Generalizations of the Kelly Criterion, a (Weinberger, 2009)	0
adjacent	2007	The value of information in financial markets: An agent-based simulati (Toth & Scalas, 2007)	0
adjacent	2019	Hawkes processes for credit indices time series analysis: How random a (Bahamou et al., 2019)	0
adjacent	2026	Learning Decision-Sufficient Representations for Linear Optimization (Ye et al., 2026)	0
adjacent	2014	Market Making with Decreasing Utility for Information (Dudik et al., 2014)	0
adjacent	2022	The Information Bottleneck Principle in Corporate Hierarchies (Gordon, 2022)	0
adjacent	2026	Optimal Rates for Agentic Networked Information Aggregation (Bateni et al., 2026)	0
adjacent	2019	Conservation Laws in a Limit Order Book (Rosenzweig, 2019)	0
adjacent	2021	The Homogenous Properties of Automated Market Makers (Jensen et al., 2021)	0
adjacent	2026	Evidence Sufficiency Under Delayed Ground Truth: Proxy Monitoring for (Solozobov, 2026)	0
adjacent	2026	Support sufficiency as action-sufficient compression: a single-cycle r (Walsh, 2026)	0
adjacent	2013	Abstraction in decision-makers with limited information processing cap (Genewein & Braun, 2013)	0
adjacent	2023	Financial Time Series Forecasting using CNN and Transformer (Zeng et al., 2023)	0
adjacent	2026	Monodense Deep Neural Model for Determining Item Price Elasticity (Garg et al., 2026)	0
adjacent	2026	The Sufficiency-Conciseness Trade-off in LLM Self-Explanation from an (Zahedzadeh & Bahrak, 2026)	0
adjacent	2025	ASTER: Adaptive Spatio-Temporal Early Decision Model for Dynamic Resou (Chen et al., 2025)	0
adjacent	2022	Task-Induced Representation Learning (Yamada et al., 2022)	0