When Should an AI Agent Delegate Execution? Counterfactual-Stable Route Certification in Intent-Based Markets

Anonymous Author(s)

Abstract

An autonomous agent holding an already-approved trading intent must choose, at creation time, whether to delegate execution to a solver market, execute directly against an automated market maker, or abstain from certifying either route. Existing on-chain execution-quality evaluation measures average realized quality for *completed* swaps against a settlement-relative baseline. That answers a different question: it scores outcomes after the fact rather than certifying a decision made before the outcome is observable, and it conditions away intents that never complete.

This paper specifies the estimand, the certification rule, and the leakage-controlled retrospective protocol for the creation-time decision, and states the fail-closed conditions under which it may be estimated. Payoffs are identified as net quantities of the buy token. The direct counterfactual is a deterministic block-start envelope over the first three canonical blocks after intent creation, restricted to officially deployed pools and required to satisfy the signed minimum-buy constraint. An unfilled delegated intent carries a real payoff — retained sell inventory valued at a prespecified block-start mark — rather than a fabricated solver loss. A route is certified only when the entire envelope clears a frozen threshold in one direction; sign instability across feasible execution blocks yields *Abstain* by construction.

This is a specification and protocol contribution supported by engineering validation evidence. We report pipeline validation and pilot feasibility diagnostics, but make **no scientific claim about comparative route performance**. Resolving data-adequacy before model comparison is the methodological point: a negative feasibility outcome is a valid protocol outcome, and a failed gate must stop downstream route-performance claims rather than trigger a substitute analysis.

CCS Concepts

• **Applied computing** → **Economics**; • **Computing methodologies** → *Machine learning*.

Keywords

intent-based markets, execution certification, selective prediction, partial identification, counterfactual stability, preregistration

1 Introduction

An agent holds a signed, already-approved intent to sell a fixed quantity of one token for another. It must choose one of three actions at creation time: *Delegate* the intent to a solver market, execute *Direct* against an automated market maker, or *Abstain* from certifying either route. The choice is made before any settlement is observed.

The existing evaluation literature answers a different question. On-chain execution-quality studies report average realized quality for orders that completed, benchmarked against a baseline constructed at or near settlement. Three gaps separate that measurement from the decision an agent actually faces.

The counterfactual reference point. A settlement-relative baseline uses information unavailable when the agent commits. To certify a decision, the direct baseline must be constructed from the block-start state of candidate execution blocks the agent could actually reach from the decision time.

Completed executions only. Restricting analysis to filled orders conditions on the outcome. An intent that expires unfilled is not missing data and not evidence of solver failure; it is a delegated non-completion that must carry a defined payoff.

Label stability. If the sign of the route advantage flips across feasible execution blocks, then a point estimate computed at one chosen block is not a decision-grade quantity. Stability must be measured, not assumed.

Contributions

This paper contributes a specification and a protocol. It does not report comparative route-performance findings.

(1) **An identified estimand.** Both action payoffs are net quantities of the buy token, never a mixture of token units. The direct counterfactual is a deterministic envelope over three consecutive post-creation blocks under a block-start convention. Unfilled delegation is valued at a prespecified retained-inventory mark.
(2) **A certification rule with abstention as a first-class outcome.** A route is certified only when the whole envelope clears a frozen threshold in one direction. Mixed signs, ties, and inadmissible endpoints all yield *Abstain*, which converts label fragility into a withheld claim rather than a coin flip.
(3) **A leakage-controlled retrospective protocol.** Rolling-origin cohorts over an Active-Trader panel, an information clock admitting a feature only when it was available strictly before intent creation, and a sealed chronological holdout whose identifier list is hashed before any outcome is opened, with no resealing path.
(4) **A fail-closed adequacy gate.** Thirteen named checks must pass before any result may be claimed. Data adequacy is resolved before model comparison, so a negative outcome remains a valid result.

What this paper is not

No economic route-performance results are reported. Every economic threshold and risk budget remains unfrozen pending a reviewed freeze procedure. The strongest claim the protocol can eventually support is a *retrospective, strictly time-ordered* model evaluation on an Active-Trader Intent Panel. It is not prospective confirmation, not a causal execution benefit, not deployed performance, and not population-representative for all users of the studied protocol.

Overview

Figure 1 separates the two halves of the design that the rest of the paper keeps separate. The upper half is the creation-time decision: the signed intent and the strictly-prior block state are the only admissible inputs, offline-trained bound models with a split-conformal residual correction produce a predicted premium interval $[\underline{\hat{\pi}}_i, \overline{\hat{\pi}}_i]$, and the three-way rule certifies a route only when that interval clears the frozen threshold in one direction. The lower half is the retrospective machinery that never touches decision time: the deterministic block-start direct counterfactual over B_{i1}, B_{i2}, B_{i3}, the delegated payoff that values an expired intent at a retained-inventory mark rather than a fabricated solver loss, the resulting true premium interval $[\underline{\pi}_i, \overline{\pi}_i]$, and the fail-closed adequacy gate that decides whether the sealed holdout may be opened at all. The single dashed arrow carries the direction of permitted information flow: retrospective labels train and calibrate the models, and nothing flows back from a label into a decision-time feature, a boundary, a threshold, or a hyperparameter.

2 Related work and positioning

Intent-based markets delegate execution to competing solvers that search for feasible ways to satisfy user-specified constraints. Existing work primarily studies the strategic behavior, entry incentives, and auction structure of these markets. Chitra et al. [5] model costly solver entry and search effort, showing that congestion and fixed participation costs can produce restricted entry. Canidio and Henneke [4] study the efficiency and fairness properties of combinatorial trade-intent auctions in which no numeraire is available to redistribute batching gains, and propose an endogenous best-execution benchmark. CoW Protocol operationalizes this market structure through combinatorial batch auctions that combine user-to-user matching with external on-chain liquidity [6]. Our question is complementary: rather than characterizing solver equilibrium or auction design, we ask whether an execution agent can certify, at intent creation, that delegation or direct execution against an automated market maker is preferable under block-level counterfactual uncertainty and possible delegated non-completion. The direct counterfactual depends on concentrated liquidity, multiple fee tiers, and per-pool state, which we compute against the deployed specification [1].

On-chain execution-quality measurement reports realized quality for completed swaps against historical baselines. This is an ex-post outcome measurement, not an ex-ante decision certification, and it typically excludes non-completing orders.

Research on transaction ordering and extractable value explains why a counterfactual execution reference is fragile: inclusion and ordering are contested [7, 11]. That literature targets the magnitude of extraction rather than pre-trade route certification, but it motivates treating a single-block counterfactual as an unstable target.

Selective prediction and conformal risk control provide the machinery for withholding a prediction under uncertainty [2, 8, 9, 12]. Neither abstention nor conformal calibration is claimed as novel here. The contribution is binding them to a two-route *economic* certification in which *Abstain* carries a defined payoff and the route label may be unidentified.

Partial identification is the closest methodological relative [10]: the object of interest is a set rather than a point. We report envelope bounds instead of collapsing the target to one block's realization.

Clustered inference with few effective clusters requires bootstrap corrections rather than naive standard errors [3], which the statistical design below adopts because order rows are not independent replicates.

3 Estimand and decision contract

3.1 Unit and primary sample

The unit is one already-approved, sell-kind, fill-or-kill intent on the studied mainnet for a single canonical token pair in either direction. The primary sample is machine-locked to that chain and pair, non-partially-fillable orders, explicit application-data market evidence, standard token balance sources, and empty pre/post interactions. On-chain placement and wrapped-native flow orders are outside the primary sample.

Fulfilled and *Expired* are the primary terminal states. *Cancelled* is a competing user override, reported separately and excluded from the primary terminal target because cancellation timing and user intent change are not consistently identified from public records.

For intent i:

- s_i, b_i: sell and buy tokens.
- q_i = `sellAmount` + `feeAmount`: the frozen total sell-token budget in base units, normalized only after token decimals and fee semantics are verified.
- a_i: the signed minimum buy quantity the execution is permitted to deliver for the frozen full sell amount.
- t_i: the public order-book creation timestamp, interpreted as the decision time.
- v_i: the intent expiry (`validTo`).
- $C_i = 1$ when a complete fill is publicly evidenced no later than v_i; $C_i = 0$ for an uncancelled expired order.

Every payoff is a net quantity of b_i. For a primary fulfilled order, `executedSellAmount` must equal q_i; a mismatch is an integrity exclusion, not a low outcome. `executedBuyAmount` is the verified receiver output and is never reduced a second time by the executed fee.

3.2 Information clock

The policy acts at t_i. A feature is admissible only when its recorded availability time is no later than t_i.

- Signed intent fields are available at creation.

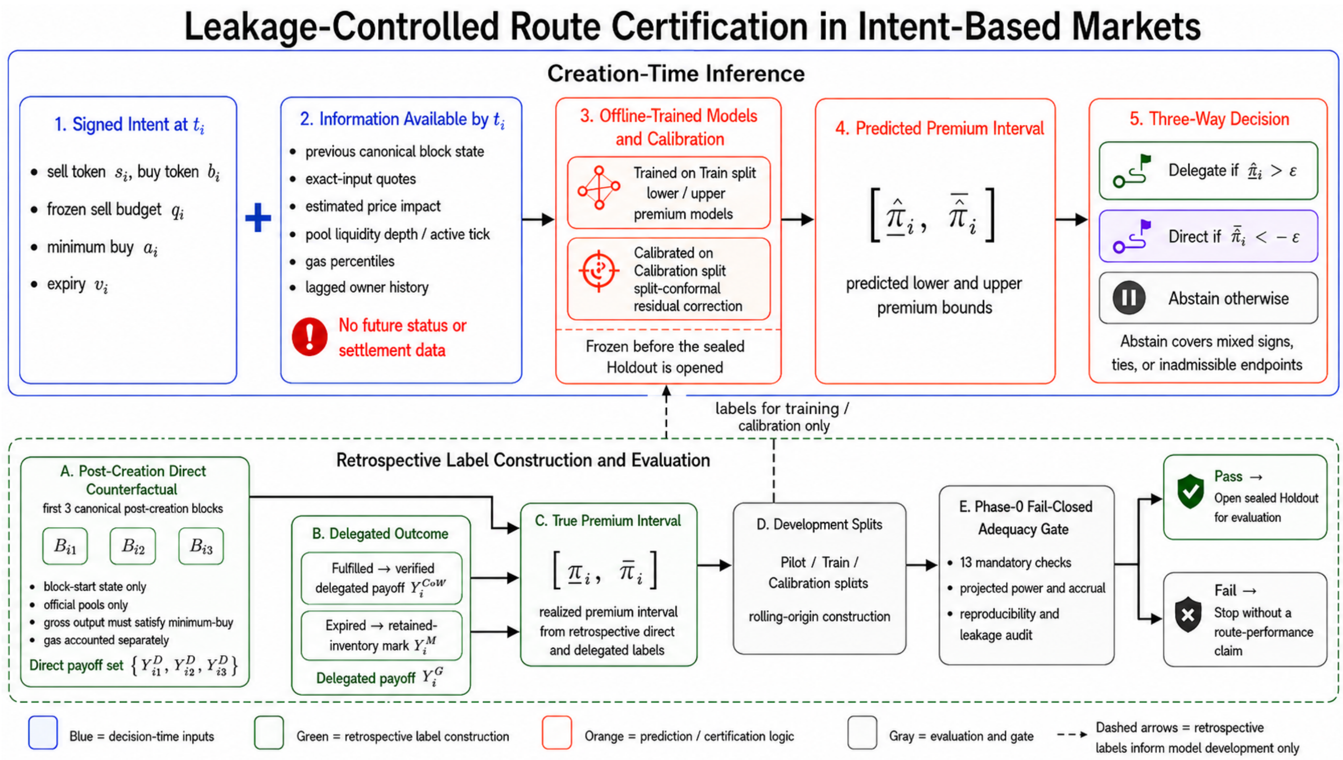

Figure 1: Leakage-controlled route certification. Top (creation-time inference): admissible inputs at t_i are the signed intent fields and state from the last canonical block strictly earlier than t_i; no future status or settlement data enters. Bound models trained on Train and calibrated on Calibration are frozen before the sealed holdout is opened, and emit a premium interval that the three-way rule maps to *Delegate, Direct*, or *Abstain*. Bottom (retrospective label construction and evaluation): the direct payoff set comes from a block-start envelope over the first three canonical post-creation blocks restricted to official pools and the signed minimum-buy constraint, with gas accounted separately; the delegated payoff is the verified fill when fulfilled and the retained-inventory mark when expired. Development splits are built by rolling origin, and the Phase-0 fail-closed gate must pass all thirteen checks before the sealed holdout is opened — otherwise the protocol stops without a route-performance claim. Dashed arrow: retrospective labels inform model development only and never become features.

- On-chain state features use the end state of the last canonical block whose timestamp is *strictly* earlier than t_i. Timestamp equality is treated as ambiguous and cannot be used.
- A lagged completion becomes available only after its settlement event; an expiry only after its v_i. A current API status does not backdate availability.

Forbidden as features: the current order's future status, executed amounts, settlement block, winning solver, auction competition state, historical quotes, realized envelope state, expiry-time state, and any future aggregate.

Counterfactual outcomes may use later public history because they are *labels*. They may not feed back into decision-time features, owner selection, calendar boundaries, threshold selection, or hyperparameter search.

3.3 Deterministic direct envelope

Let B_{i1}, B_{i2}, B_{i3} be the first three consecutive canonical blocks with timestamps strictly greater than t_i. The simulated swap in block B_{ik} starts from the end state of its parent block. This block-start convention avoids using transactions that occurred earlier within the execution block itself.

Each baseline additionally requires the execution-block timestamp to be no later than v_i. A baseline past expiry is marked time-authorization-failed and excluded even if its price would be favorable. Other non-executable routes are marked `DIRECT_UNAVAILABLE`. Blocks are never skipped until a favorable price appears.

The candidate set is every pool emitted by the official factory for the exact token pair at the frozen fee tiers, provided the pool was initialized and had executable block-start liquidity. The primary route is one single pool, not a split route or an aggregator. For each block and fee tier, the identical q_i is simulated. A pool is executable only if its gross receiver output satisfies $O_{ikp}(q_i) \geq a_i$; gas is accounted separately and never used to waive the signed token constraint.

For candidate pool p and baseline k:

$$Y_{ik}^{D,p} = O_{ikp}(q_i) - G_{ikp}, \qquad p_{ik}^* = \arg\max_p Y_{ik}^{D,p}, \qquad Y_{ik}^D = Y_{ik}^{D,p_{ik}^*}.$$

Here O is exact-input pool output and G is direct user gas converted to b_i at the block-start reference price. Ties are resolved by lower fee tier, then lower checksummed pool address. The rule is applied without reference to the delegated outcome.

If no candidate pool satisfies the signed minimum in any of the three baselines, the policy performs a deterministic block-start preflight and does not broadcast an invalid swap, so attempted-swap gas is zero. Its utility is then the block-start marginal mark of the retained q_i, with execution indicator $E^D_{ik} = 0$.

3.4 Retained-inventory mark

For each official pool initialized with positive active liquidity at block start, let $r_{ip}(H)$ be the instantaneous buy-token output per unit sell token implied by the pool's price state at block H. The marked quantity is

$$M_{ip}(H) = q_i \left(1 - \frac{f_p}{10^6}\right) r_{ip}(H), \qquad M_i(H) = \max_p M_{ip}(H).$$

The mark applies the pool fee but no finite-size price impact and no gas. It ignores the signed minimum because it is a valuation, not a broadcast. If no eligible official pool exists, the mark is `VALUATION_UNAVAILABLE`.

The same rule is used for unavailable direct execution, expired-intent inventory valuation, *Abstain*, and reference normalization.

3.5 Gas scenarios

Baselines $D1$ and $D2$ use median gas units and median priority fee. Baseline $D3$ uses the 95th-percentile gas units and 90th-percentile priority fee. Every quantile uses the nearest-rank rule $\lceil qn \rceil$ on frozen development receipts.

The cell hierarchy is fixed: direction × fee tier × tick-crossing bucket, then direction × fee tier, then direction, then pooled receipts. A cell with fewer than 30 receipts falls deterministically to the next broader cell. If the pooled cell is also short, gas is `GAS_UNAVAILABLE`, the row is forced to non-certifying *Abstain*, and no statistic is chosen ad hoc.

The operational *Direct* action executes $D1$. $D2$ and $D3$ are latency/gas perturbations used to test whether certification is counterfactually stable. They are not three actions from which an ex-post policy may cherry-pick. Because $D3$ moves both block timing and gas conservatism, it is a joint stress rather than an identified latency effect; a required sensitivity analysis decomposes the two axes without replacing the primary envelope.

3.6 Delegation payoff

The primary *Delegate* policy is a single complete delegated policy: submit the approved intent, wait until it is fulfilled or reaches v_i, and if it expires unfilled, retain the sell inventory and value it at the prespecified block-start mark after v_i. Its end-to-end payoff is

$$Y_i^G = \begin{cases} Y_i^{\mathrm{CoW}}, & C_i = 1, \\ Y_i^M, & C_i = 0, \end{cases}$$

where Y_i^{CoW} is the verified net buy-token quantity credited to the order receiver and Y_i^M is the retained-inventory mark. Waiting cost is represented by the later mark, not by inventing a solver loss. Expired means delegated non-completion. A fulfilled transfer below the signed minimum is an integrity failure, not a low outcome.

3.7 Decision rule

Both payoffs are normalized by a common reference quantity so the premium is scale-free and unit-consistent. The premium is defined only when both endpoints are admissible; when either is unavailable it remains undefined and contributes to coverage reporting rather than to a numeric estimate.

Let $\underline{\pi}_i$ and $\overline{\pi}_i$ be the lower and upper bounds of the normalized premium over the three baselines, and let ε be the frozen certification threshold. Then

$$\mathrm{action}_i = \begin{cases} \textit{Delegate}, & \underline{\pi}_i > \varepsilon, \\ \textit{Direct}, & \overline{\pi}_i < -\varepsilon, \\ \textit{Abstain}, & \text{otherwise}. \end{cases}$$

Abstain covers mixed signs, ties, and every inadmissible endpoint. It is not free: it is scored under a fixed prespecified default. Sign instability across the envelope therefore yields *Abstain* by construction.

ε is prespecified from an economically meaningful SESOI and frozen before the sealed holdout is opened. It is not tuned on Train, Calibration, or Holdout data and is never chosen to make a route win.

4 Retrospective evaluation protocol

4.1 Rolling-origin cohorts

Each cohort is a discovery window followed by an evaluation window at a fixed step. Owner eligibility requires a public settlement event strictly before the evaluation window opens, which prevents an intent's own later settlement from selecting its owner into the cohort that evaluates that intent. An intent is evaluated only when its creation time falls in the left-closed, right-open evaluation interval. When evaluation windows overlap, an identifier is assigned to the earliest eligible cohort only.

Cohorts nest inside macro chronological partitions: Pilot, Train, Calibration, and a sealed Retrospective holdout. Macro boundaries are selected from coverage metadata only. No label, premium, stability value, loss, or model output may choose or move a boundary. The holdout identifier ledger is hashed before any outcome is opened, and there is no resealing path: if a candidate holdout outcome was already opened, the contract fails rather than selecting a different period.

4.2 Panel scope and its limit

The design is an *Active-Trader* panel, conditional on prior successful settlement. Public settlement logs discover completed orders and their owners, and public owner history recovers later fulfilled, expired, and cancelled orders for those owners. Public sources cannot enumerate never-settling users or establish an unbiased frame for all intents. Panel rates therefore may not be transported to the full user population.

Owner-history results are current terminal records, not historical snapshots. Cancellation timing, complete competition records, and permissioned auction state are not reconstructed.

4.3 Readiness diagnostics and audit

Readiness targets over terminal intents, expired intents, and distinct expired owners are *planning diagnostics*. Meeting them reports `descriptive_targets_met=true`; it never creates a scientific pass and never freezes a model, threshold, test period, or claim. Pilot identifiers are counted as development observations and removed before the targets are evaluated.

The offline audit reports, globally and per cohort: terminal-state counts; distinct expired owners; owner and date concentration, including effective cluster counts; direction counts; duplicate identifiers; discovery-before-evaluation violations; orders outside their declared window; invalid or naive timestamps, which are never silently coerced; and source and artifact hashes. Any integrity violation sets `descriptive_targets_met=false` even when raw counts exceed the targets.

4.4 Inference

The primary 95% interval is the convex hull of three clustered constructions: an owner-cluster wild bootstrap-t interval, a seven-day moving-block interval, and a two-way owner-by-week interval. Small-cluster support gates apply and order rows are never treated as independent replicates [3].

Premium, completion, and selective-risk endpoints have separate effect-size grids and power calculations. A projected minimum detectable effect is computed *before* the holdout is opened, using Train and Calibration cluster structure and ex-ante candidate counts. Post-hoc observed power is forbidden. Power simulation runs the complete pipeline — sampling, fitting, calibration of prediction bounds and application of the prespecified threshold, matched coverage, abstention, rare-expiry handling, and recalibration — because raw order count is not the effective sample size.

Model complexity is capped by design: simple decisive baselines, a small fixed model set, and an ex-post oracle. A broad model menagerie is excluded so that the evaluation surface rather than architecture search carries the contribution.

4.5 Decision model and calibration

The predictive models are fitted exclusively on Train data. Conformal residual corrections and calibration parameters are fitted exclusively on Calibration data. All components, including ε, are frozen before the sealed holdout is opened. The intended pipeline (to be reviewed at protocol freeze) corresponds to panels 2–5 of Figure 1 and is:

(1) **Feature construction.** All market-state features are constructed from the end state of the last canonical block strictly preceding t_i, while signed intent features $(q_i, a_i, v_i - t_i)$ are observed at t_i. Admissible features include: current exact-input quotes and estimated price impact at that block; pool liquidity depth and active tick; frozen gas percentiles (base fee and priority fee from the development window); lagged owner-history statistics; and direction/size indicators. The realized post-creation envelope, its width $(\overline{\pi}_i - \underline{\pi}_i)$, completion status, premium, and expiry-time mark are *labels only* and are never used as features.

(2) **Envelope estimation.** Separate gradient-boosted regression or quantile-regression models are fitted on Train data: one targets $\hat{\underline{\pi}}_i$ (lower bound) and one targets $\hat{\overline{\pi}}_i$ (upper bound). This is a regression problem, not classification.

(3) **Calibration.** Split-conformal calibration on the Calibration set provides one-sided residual corrections for each predicted bound, yielding nominal marginal-coverage intervals whose realized coverage is audited under the rolling-origin protocol [2, 12]. Because the data are temporally ordered and owner-clustered, no unrestricted distribution-shift guarantee is claimed. Threshold ε is set to the prespecified SESOI (e.g., 10 bps) based on economic significance; it is not adjusted to achieve a target abstention rate and is never selected on the sealed holdout.

(4) **Three-way action.** At test time: if $\hat{\underline{\pi}}_i > \varepsilon$, emit *Delegate*; if $\hat{\overline{\pi}}_i < -\varepsilon$, emit *Direct*; otherwise emit *Abstain*. The abstention rate and risk–coverage frontier are primary endpoints alongside the route premium.

The exact model family and calibration procedure are not frozen until the reviewed protocol freeze; Phase 0 asks only whether the sample size and cluster structure can support this pipeline.

4.6 Fail-closed adequacy gate

Before any result may be claimed, thirteen named checks must pass. In summary they require: validated scope and authorization with downstream gates closed; recomputation of every check by its allowlisted deterministic verifier in an isolated clean worktree, matching command, exit status, canonical output digest, and source hashes; bounded public retrieval with provenance replay; mechanical boundary resolution without outcome-based date selection; lagged-only owner and pagination semantics; raw-cache replay and provenance sealing with quarantine of opened identifiers and no reseal path; reconstruction accuracy and route-reversal thresholds; projected coverage and effective-cluster support; projected clustered power; leakage and contamination ledgers in which access time never substitutes for event time; absence of paid data, permissioned endpoints, and any prospective collector; two independent reviews reporting zero correctness blockers; and deterministic artifact replay.

The gate fails closed. A boolean assertion receipt, a self-hash, or review prose containing pass markers is never decision authority. Before any evidence run, a reviewed lock pins the exact validator blob, runtime fingerprint, dependency-lock hashes, verifier command manifest, and verifier blobs at a lock commit; later evidence commits must prove those blobs unchanged.

The gate asks whether the question can be estimated precisely, not which route wins. A negative economic result is valid.

5 Results

We report three categories of pre-gate evidence: Uniswap V3 reconstruction validation, a pilot engineering summary, and a power and accrual projection. These are engineering-validation and feasibility results, not comparative route-performance findings. The Phase-0 feasibility gate yielded `FEASIBILITY_FAIL` on accrual (Section 5.3), so no sealed holdout exists and no scientific claim about route premiums, completion rates, or selective-risk trade-offs may be drawn from this data.

5.1 Uniswap V3 Reconstruction Validation

We validated the swap-output reconstruction component of the direct-envelope pipeline against 562 real on-chain WETH/USDC swap transactions (candidates identified by stratified sampling across size quartiles and both directions) over a 100,000-block window (blocks 25,461,362–25,561,362). Of the 562 candidates, 81 were verified by Method A (on-chain quoter replay) and an independent set of 101 by Method B (Settlement-event state replay), for 182 total verified transactions. The simulated `amountOut` was compared against the on-chain record for each verified swap.

Table 1: Uniswap V3 reconstruction validation (engineering pilot). All comparisons against on-chain `amountOut`.

Metric	Method A (quoter)	Method B (event replay)
Swaps checked	81	101
Exact match ($\Delta = 0$)	81	101
Nonzero errors	0	0
Label flips possible	0	0
Prefix challenges passed	20	

Every reconstructed output matched the on-chain record exactly (182/182 = 100%). No reconstruction error was large enough to change the route-advantage sign for any non-tied order. This validates that the swap-output reconstruction component of the pipeline correctly recovers historical `amountOut` from public event data alone, without paid archive access. Gas estimation, route selection, and the full counterfactual envelope remain subject to the protocol freeze review.

5.2 Pilot Engineering Summary

Classification: DEVELOPMENT-ONLY PILOT. Non-scientific. Not a feasibility claim. Window: 2026-07-15 to 2026-07-18 (3.2 days). All rows marked `final_test_eligible=false`.

Table 2: Pilot engineering summary (3.2-day window, development evidence only). Figures are engineering diagnostics, not scientific estimates.

Item	Count / Rate
Orders evaluated	68
Fulfilled	40
Expired (non-completion)	28
WETH→USDC / USDC→WETH	42 / 26
Fulfilled, all 3 baselines admissible	17 (42.5%)
Fulfilled, $\geq$ 1 baseline inadmissible	23 (57.5%)
Gas-infeasible D1	13
Limit-infeasible D1	16
Comparable: solver-favored (robust)	14 / 17
Comparable: direct-favored (robust)	1 / 17
Comparable: abstain	2 / 17
Median robust premium	4.2 bps
Median direct envelope width (D1–D3)	1.8 bps
Sign-instability rate (CSI, abstain)	11.8% (2/17)
Distinct comparable owners	12

Two findings are methodologically informative independent of their numeric values. First, 57.5% of fulfilled orders had at least one baseline inadmissible — primarily because the assumed 180,000-gas budget rendered direct execution economically inadmissible after gas accounting. This confirms that the `GAS_UNAVAILABLE` and `INFEASIBLE_LIMIT` codes are not edge cases but routine outcomes of the certification filter, and that comparison coverage must be tracked as a primary endpoint. Second, 28 expired orders were recovered as non-completion labels with defined retained-inventory payoffs, confirming that the pipeline correctly handles the non-completed-delegation case without inventing solver-loss values.

The 3.2-day window is too short to support any scientific claim. Observed quantities carry wide pilot uncertainty: the 95% Wilson interval on comparison coverage is [0.29, 0.57].

5.3 Power and Accrual Gate

Table 3: Required comparable sample sizes for the robust-premium endpoint at 90% power ($\alpha = 0.05$, two-sided), derived from pilot σ. The bold cell denotes the primary design candidate.

SESOI	DEFF 1	DEFF 1.5	**DEFF 2**	DEFF 3	
5 bps	131	196	261	392	σ = 17.5 bps
10 bps	35	52	**69**	103	
20 bps	11	16	21	31	

(winsorized pilot estimate). Raw pilot σ = 36.3 bps; the primary design candidate uses 10 bps SESOI, DEFF 2, requiring 69 comparable orders (winsorized) or 281 (raw).

The observed comparable-order rate from Table 2 is 17 comparables / 3.2 days $\approx$ 5.3/day. A 12-day near-term projection gives $5.3 \times 12 \approx 64$ comparable orders — below both the winsorized target (69) and the conservative target (281). Meeting 281 comparable orders at the observed rate would require $281/5.3 \approx 53$ calendar days.

The Phase-0 feasibility gate returned `FEASIBILITY_FAIL` on the `accrual` check. All other mandatory checks passed: scope, CoW pipeline, Uniswap reconstruction (exact match: 182/182 = 100%), counterfactual signal pipeline, and reproducibility. The infrastructure is validated; the failure is a calendar constraint, not a methodological defect. This is the expected outcome of a fail-closed design: the gate prevents underpowered scientific claims rather than producing them.

6 Limitations

Selection. The panel is conditional on prior successful settlement. Estimates apply to that frame, not to all users of the protocol.

Market-side counterfactual. The direct envelope is a market-side route, price, liquidity, and gas object. It is conditional on sufficient sell-token balance, allowance or permit, valid nonce and signing authority, and permission to use the receiver and route. Those wallet facts are not reconstructed from an event-only panel, so the envelope may not be described as unconditional wallet-level executability. Any live deployment must re-check them.

Retrospective evidence. Every market observation predates the protocol freeze, and engineering pilot outcomes were inspected beforehand. The evidence is retrospective and strictly time-ordered; it is not prospective confirmation. Pilot observations extend past the mechanical candidate horizon and are therefore recorded as later external development evidence, admissible for engineering validation but not for features, thresholds, model choices, effect sizes, or claims.

Label semantics. Expired means delegated non-completion, not solver failure. Cancelled orders are excluded from the primary terminal target because cancellation timing and user intent change are not consistently identified.

Ordering and extraction. Adversarial inclusion and ordering are not fully reconstructed. The envelope is a deterministic block-start simulation, not a guarantee about contested inclusion [7, 11].

Component novelty. Abstention, gradient boosting, and conformal calibration are not individually novel. The contribution is the identification contract that binds them to a two-route economic certification.

7 Reproducibility

Every source is public: settlement events, the public owner-orders endpoint, and official factory and pool events with block, receipt, and gas evidence. No paid dataset, paid archive access, quote subscription, or permissioned endpoint is required. Rate limits affect engineering throughput but never the sample definition; a bounded failure is reported rather than silently substituting a paid source.

Acquisition is a bounded, explicitly invoked, resumable backfill — not a daemon, scheduler, or continuously running collector. Every artifact is content-addressed. The coverage ledger requires raw rows to equal emitted rows plus explicitly classified exclusions. The builder independently recomputes row identities from raw caches and requires equality among cache replay, seal ledger, and materialized tables, so recomputing a hash after altering a timestamp, identifier, status, amount, token, owner, or provenance field does not make a forged row admissible.

8 Conclusion

We specify the estimand, certification rule, and retrospective protocol for the creation-time choice among delegation, direct execution, and abstention, and we state the fail-closed conditions under which it may be estimated. Payoffs are identified as net buy-token quantities; the direct counterfactual is a decision-time block-start envelope; unfilled delegation carries a real retained-inventory payoff; and counterfactual instability produces abstention rather than a coin-flip label. No route-performance claim is made here: the infrastructure is validated, pilot evidence is development-only, and a negative feasibility outcome is an equally valid result.

References

[1] Hayden Adams, Noah Zinsmeister, Moody Salem, River Keefer, and Dan Robinson. 2021. *Uniswap v3 Core.* Technical whitepaper. Uniswap Labs.

[2] Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning* 16, 4 (2023), 494–591.

[3] A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. 2008. Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics* 90, 3 (2008), 414–427.

[4] Andrea Canidio and Felix Henneke. 2024. Fair Combinatorial Auctions: Endogenous Best Execution in Blockchain Trade-Intent Markets. arXiv preprint arXiv:2408.12225.

[5] Tarun Chitra, Kshitij Kulkarni, Mallesh Pai, and Theo Diamandis. 2024. An Analysis of Intent-Based Markets. arXiv preprint arXiv:2403.02525.

[6] CoW Protocol. 2024. CoW Protocol Documentation: Batch Auctions and Solver Competition. Protocol documentation.

[7] Philip Daian, Steven Goldfeder, Tyler Kell, Yunqi Li, Xueyuan Zhao, Iddo Bentov, Lorenz Breidenbach, and Ari Juels. 2020. Flash Boys 2.0: Frontrunning in Decentralized Exchanges, Miner Extractable Value, and Consensus Instability. In *IEEE Symposium on Security and Privacy (SP).* IEEE, San Francisco, CA, USA, 910–927.

[8] Ran El-Yaniv and Yair Wiener. 2010. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research* 11, 53 (2010), 1605–1641.

[9] Yonatan Geifman and Ran El-Yaniv. 2019. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *International Conference on Machine Learning (ICML) (Proceedings of Machine Learning Research, Vol. 97).* PMLR, Long Beach, CA, USA, 2151–2159.

[10] Charles F. Manski. 2003. *Partial Identification of Probability Distributions.* Springer, New York.

[11] Kaihua Qin, Liyi Zhou, and Arthur Gervais. 2022. Quantifying Blockchain Extractable Value: How dark is the forest?. In *IEEE Symposium on Security and Privacy (SP).* IEEE, San Francisco, CA, USA, 198–214.

[12] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World.* Springer, New York.