Fill Probability Is Not Enough: Empirical Fill Modeling and Net-Outcome Prediction for BTC/USDT Market Making

Michel Bassil Chandrasekhara Devarakonda

April 2026

Abstract

This paper asks whether short-horizon fill prediction is enough to make passive BTC/USDT market making work. Using Binance 100 ms L2 book updates and tick-by-tick trades from 21 trading days, we first fit empirical fill curves and then test several quoting objectives in a market-making simulator with fees, slippage, adverse selection, inventory limits, and end-of-day liquidation. On the prediction side, the problem looks clean: an exponential form explains more than 99.9% of the cross-depth variation in fill rates, and feature-based classifiers reach out-of-sample AUC near 0.87. Turning those predictions into PnL is harder. Strategies that rank quotes mainly by fill probability lose money across all intraday volatility buckets, even after adding gating or a separate adverse-selection model. The first consistently positive aggregate result appears when the model is trained directly on realized order-level net outcome, with non-filled orders labeled as zero. Adding longer-context features improves selectivity further and raises single-level performance to +$21.8 over 21 days with annualized Sharpe about 1.03, while multi-level quoting stays negative. The overall lesson is simple: in this medium-frequency setting, predicting fills well is useful, but it is not the same thing as predicting good trades.

1 Introduction

Passive market making starts with execution. A trader posts bid and ask limit orders, earns the spread only if those orders trade, and can easily lose that spread again when fills arrive just as the market moves the wrong way. The trade-off is immediate: quotes closer to the midprice fill more often but earn less per trade, while deeper quotes earn more spread when they do fill but sit unexecuted much more often.

That naturally leads to the standard question: if a trader posts a limit order at depth d from the midprice, what is the chance that it fills within horizon T? Fill probability matters, but it is not the quantity that a trading system ultimately cares about. What matters in deployment is the net result after fees, slippage, adverse selection, inventory build-up, and end-of-day liquidation.

This paper does three things. First, it documents a stable empirical fill structure for BTC/USDT using only public data at medium-frequency resolution. Second, it shows that good fill prediction does not automatically translate into profitable passive market making once realistic costs are included. Third, it finds that direct prediction of order-level net outcome works better than proxy objectives that model fill probability and adverse selection separately.

2 Data and Synthetic Order Construction

Market data

The dataset contains BTC/USDT Binance spot order-book updates at 100 ms L2 cadence and tick-by-tick trades. L2 data provide displayed depth at multiple price levels but not individual order identities. The sample covers 21 trading days: 15 development days drawn from three volatility

regimes (5 high-volatility, 5 normal-volatility, 5 low-volatility) and 6 additional out-of-sample days that were never used during initial model construction.

Depth is measured in basis points relative to the prevailing midprice,

$$M_t = \tfrac{1}{2}(B_t + A_t),$$

where B_t and A_t are the best bid and best ask. A synthetic quote at depth d is therefore a passive order placed d bps away from M_t on either side of the book. This normalization is necessary because fixed dollar offsets are not comparable across BTC price regimes.

Sampling, horizons, and labels

At one-second decision times, synthetic bid and ask orders are evaluated on the depth grid

$$\mathcal{D} = \{0.1,\ 0.25,\ 0.5,\ 1,\ 2,\ 3,\ 5,\ 7,\ 10,\ 15,\ 20,\ 30\}\ \text{bps}$$

and the horizons $T \in \{60, 300, 600\}$ seconds. Because the data are L2 rather than full L3 order-level messages, execution must be approximated from displayed queue state and subsequent trades.

For each synthetic order, *queue-ahead* is defined as the displayed same-side resting quantity already posted at the simulated order price at submission time. A fill is recorded under a conservative crossed-through rule: opposite-side traded volume must be sufficient to consume that queue-ahead and trade through the simulated price. For a bid posted at price P_i and time t_i, the fill label over horizon T is

$$\text{fill}_{i,T} = \mathbf{1}\left\{ \int_{t_i}^{t_i+T} \mathbf{1}\{\text{trade price} \le P_i\}\, v\, dt > q_i^{\text{ahead}} \right\}, \tag{1}$$

with the inequality reversed for ask orders.

This definition is stricter than a touch rule and is intentionally conservative. It is also the rule implemented in the shared fill-checking library used by all experimental stages.

3 Models and Objectives

Fill-probability models

The parametric baseline is the exponential specification

$$P_{\text{fill}}(d \mid T, s) = A(T, s) \exp\{-\kappa(T, s) d\}, \tag{2}$$

where s denotes market state. In this parameterization, A controls fill probability near the touch and κ governs how rapidly fill probability decays with depth.

We complement the parametric model with feature-based fill classifiers. The base feature set contains ten variables: depth, log-depth, side, imbalance, absolute imbalance, relative spread, top-of-book size, log top-of-book size, queue-ahead, and log queue-ahead. Top-of-book imbalance is defined as

$$I_t = \frac{Q^{L1}_{\text{bid},t} - Q^{L1}_{\text{ask},t}}{Q^{L1}_{\text{bid},t} + Q^{L1}_{\text{ask},t}}. \tag{3}$$

Short-context models expand this to 17 features by adding momentum, recent volatility, and interaction terms. Long-context models further expand to 28 features by adding 5-minute and 30-minute returns, broader volatility state, time-of-day terms, and session-level variables.

Adverse selection and economic targets

Adverse selection is the post-placement price movement against the quote. In the two-stage setup, it is modeled as a cost term $\widehat{AS}_i$ in bps, where positive values are bad for the market maker. This

gives two natural proxy scores:

$$S_i^{\text{fill}} = \hat{p}_i\,(d_i - c), \tag{4}$$

$$S_i^{\text{two}} = \hat{p}_i\,(d_i - \widehat{AS}_i - c), \tag{5}$$

where $\hat{p}_i$ is predicted fill probability and c is explicit trading cost in bps.

The unified model dispenses with this decomposition and predicts order-level net outcome directly. The training target for order i is

$$y_i = \begin{cases} \dfrac{M_{t_i+T} - P_i}{M_{t_i}} 10^4 - c, & \text{if bid order } i \text{ fills,} \\ \dfrac{P_i - M_{t_i+T}}{M_{t_i}} 10^4 - c, & \text{if ask order } i \text{ fills,} \\ 0, & \text{if order } i \text{ does not fill.} \end{cases} \tag{6}$$

Unfilled orders are labeled as zero because a quote that never executes has no direct PnL impact in the framework used here. The resulting score is simply

$$S_i^{\text{uni}} = \widehat{\mathbb{E}[y_i \mid x_i]}.$$

4 Trading Strategy and Simulator

Single-level strategy

The core trading strategy is a passive two-sided market maker. At each decision time, the model scores all candidate bid and ask quotes on the admissible depth grid. In the single-level configuration, the strategy places at most one bid and one ask: the highest-scoring admissible quote on each side. If no quote on a side looks worthwhile, the strategy simply stays out.

The orders are passive rather than directional. A bid fill increases inventory, and that position has to be worked off later by an ask fill; the same logic applies in reverse on the short side. Inventory control therefore matters even though the quoting rule itself is market neutral.

Order lifecycle and PnL accounting

The simulator enforces explicit order lifecycle states ‘IDLE‘, ‘LIVE‘, and ‘COOLDOWN‘. A side cannot submit a new order while an earlier order on that side is still live. Orders either fill according to the measured fill-time label or time out at the chosen horizon. In the final single-level simulator, each fill has size 0.01 BTC, inventory is capped at ± 0.10 BTC, and inventory skew is set to 8 bps at the position limit. Quotes shallower than 2 bps are excluded because explicit costs already eat most of the gross spread at those depths.

PnL is computed from cash flows rather than from a hand-written spread-minus-AS decomposition. A bid fill reduces cash by executed notional plus fee and slippage; an ask fill increases cash net of those costs. Daily mark-to-market is cash plus inventory valued at the current midprice, and any remaining inventory is liquidated at the end of the day with an additional 2 bps liquidation slippage. All final strategy comparisons use this accounting.

Ablations beyond the base simulator

Several execution overlays are tested as ablations rather than built into the base simulator. These include adaptive intraday volatility gating, cancellation triggers based on adverse mid-price drift and volatility spikes, two-stage adverse-selection-aware scoring, and multi-level simultaneous quoting. Multi-level quoting allows up to six live depths per side with confidence-scaled order size and a 5 BTC capital cap, but in this sample it remains unstable economically.

5 Experimental Protocol

The empirical work separates model validation from strategy validation.

Model validation. The development set contains 15 days. Within that sample, leave-one-day-out validation is used for fill classification, adverse-selection regression, and threshold selection. Fill models are evaluated with AUC, Brier score, and calibration. Survival analysis is used as a secondary check that censoring is being handled sensibly, but the main classifier comparisons rely on fixed-horizon fill labels.

Strategy validation. Backtest quality is measured by daily cash-flow PnL, fills per day, win count across days, and annualized Sharpe computed from daily strategy returns. In addition to within-sample ablations, a dedicated six-day out-of-sample test is run on days that were sourced and cached separately from the 15-day development set. Later experiments switch to leave-one-out validation across all 21 days once the focus moves from regime-labeled development to cross-day robustness.

6 Results

Fill models work well, but profitability does not follow

The fill modeling side of the problem is unusually clean. Exponential curves fit extremely well across depth, horizon, and volatility regime, with $R^2 > 0.999$ in the pooled regime fits. Regime effects are also large: on high-volatility days the fill curves are materially flatter than on low-volatility days at the same horizon, which points to wider optimal quoting distances.

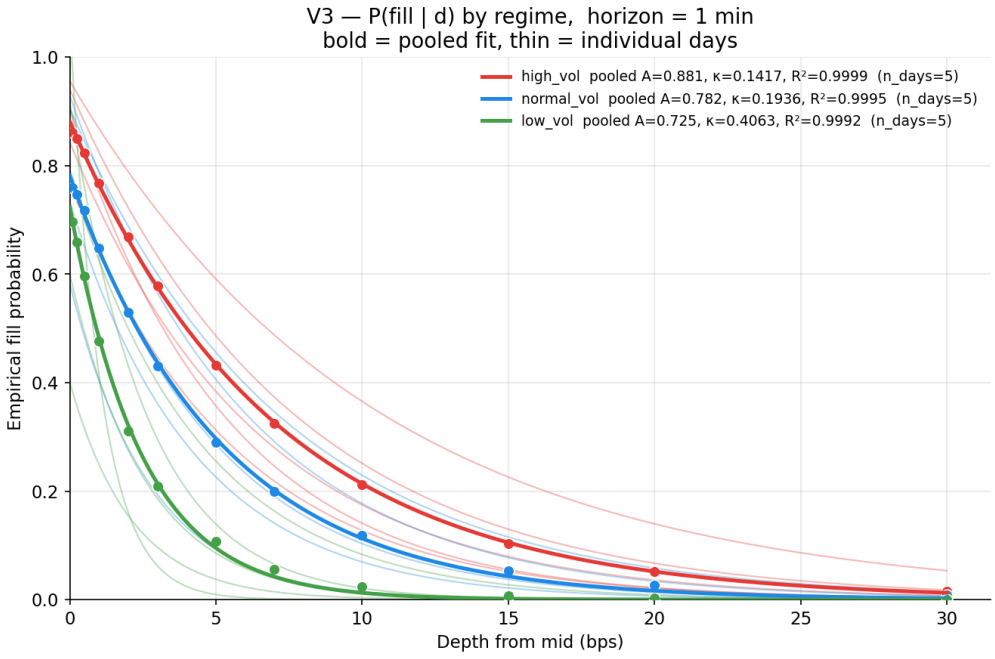

Figure 1: Empirical and fitted fill curves at the 60-second horizon by volatility regime. The separation across regimes is stable enough to justify regime-aware or state-aware execution models.

Two additional microstructure findings matter later on. First, spread-state conditioning is mostly unhelpful on BTC/USDT because the spread sits at the one-tick floor almost all the time. Second, top-of-book imbalance shows a clear mirror effect by side: bid-heavy books slow bid fills and speed up ask fills, while ask-heavy books do the opposite. Feature-based classifiers improve local discrimination beyond the exponential baseline and reach out-of-sample AUC up to about 0.87.

Even so, high AUC does not solve the economic problem. Adverse selection is fairly stable at the level of state-conditional averages but hard to predict at the individual-order level. That leaves separate fill and adverse-selection models statistically useful, but still short of what the trading problem needs.

Daily volatility regimes do not hold up out of sample

Once the simulator includes realistic order lifecycles and cash-flow PnL, the first market-making results look regime-dependent: normal-volatility development days appear profitable at the 300-second horizon, while high-volatility days remain clearly negative. That story does not survive the out-of-sample test.

Regime	Development mean PnL/day	6-day OOS mean PnL/day	Interpretation
Normal volatility	+$24.0	-$15.4	Verdict flips
Low volatility	-$10.0	+$26.7	Verdict flips
High volatility	-$53.0	-$25.1	Remains negative

Table 1: Single-level fill-based market-making results at the 300-second horizon. Two of the three regime conclusions reverse on the six true out-of-sample days, so daily regime labels are not a robust decision rule here.

This changes the direction of the project. Daily regime labels help describe fill curves, but they are too coarse to anchor a stable market-making rule. The later experiments therefore move away from day-level regime gating and focus instead on intraday volatility and direct order-level economic prediction.

Intraday volatility filters do not rescue fill-based quoting

The clearest negative result in the project is that intraday volatility selection does not rescue fill-centric passive market making. Using leave-one-out validation across all 21 days, the fill-only strategy remains negative in every realized intraday volatility bucket.

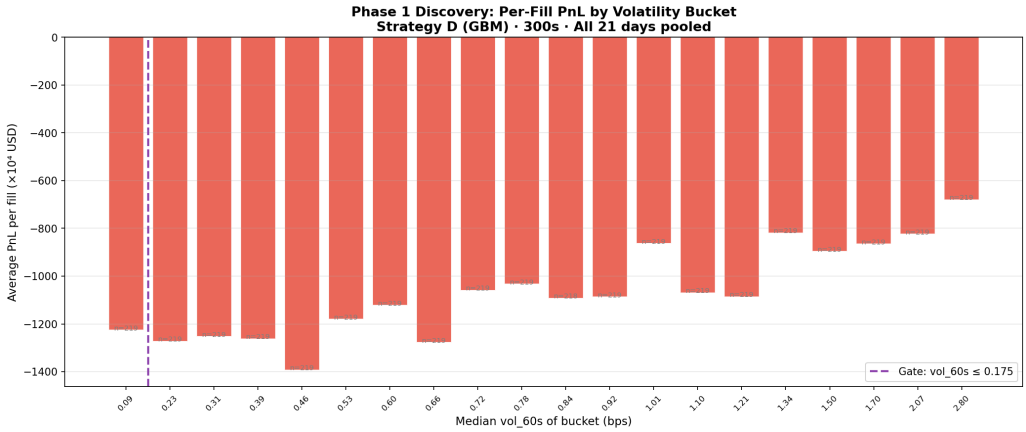

Figure 2: Per-fill PnL by intraday volatility bucket under fill-centric placement. Every bucket is negative after explicit costs; there is no obvious volatility "safe zone" for the fill-only objective.

Ungated fill-centric trading loses $827.7 over the 21-day sample. The best adaptive volatility gate cuts that loss to $300.6, but it still does not make the strategy profitable. Cancellation helps more as a damage-control mechanism: the best rule reduces total losses to $25.4, but only by shrinking activity to 2.5 fills per day on average. Put differently, cancellation works as a seatbelt, not as a way to steer the strategy into good trades.

Direct net-outcome targets produce the first positive aggregate result

Two-stage adverse-selection-aware scoring is materially better than fill-centric quoting, but it remains brittle. It works best as a gate-only ranking function and tends to over-filter when combined with aggressive score-based cancellation. One AS-aware cancellation variant does turn slightly positive, but it does so with only 0.14 fills per day on average, which is too sparse to support much of a deployment claim.

The first economically meaningful positive result appears only when the learning target is direct net outcome. The unified model scores each candidate order by predicted horizon-end net PnL, with zero targets for non-fills. Under that objective, the single-level strategy finishes the 21-day leave-one-out exercise at +$8.0 with annualized Sharpe 0.39 and 4.9 fills per day.

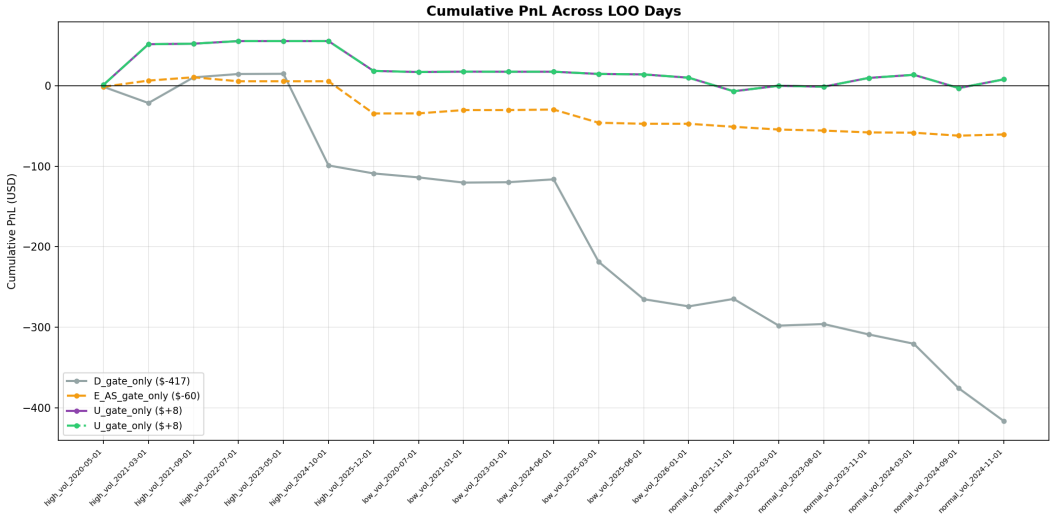

Figure 3: Cumulative 21-day PnL for fill-only, two-stage, and unified single-level scoring. Aggregate performance turns positive only when the model is trained directly on order-level net outcome.

This is the main empirical result of the paper. Fill probability matters for execution modeling, but at this data frequency and cost level it is not enough to serve as the trading objective by itself.

Longer context helps the single-level strategy, but not multi-level quoting

The final experimental stage adds 11 longer-context features to the 17-feature short-context model. These features give the model access to 5-minute and 30-minute returns, broader volatility state, time-of-day, and session-level context. The immediate effect is greater selectivity: the single-level strategy trades less often, but the trades it does take are better on average.

The best-performing configuration in the study is a conservative single-level market maker driven by a unified gradient-boosted regressor trained on the 28-feature context set to predict expected order-level net PnL at the 300-second horizon. It places at most one bid and one ask at a time, uses 0.01 BTC order size, excludes quotes shallower than 2 bps, enforces a ±0.10 BTC inventory cap with 8 bps inventory skew, charges a 1.0 bps maker fee plus 0.5 bps slippage, and liquidates any remaining inventory at the end of the day with 2 bps liquidation slippage.

Configuration	Total PnL	Fills/day	Annualized Sharpe
Fill-only, ungated	-$827.7	208.5	–
Best adaptive volatility gate	-$300.6	38.7	–
Best cancellation overlay	-$25.4	2.5	-1.31
Two-stage fill + AS, gate only	-$60.4	9.3	-4.72
Unified single-level, short context	+$8.0	4.9	+0.39
Unified single-level, long context	+$21.8	3.2	+1.03
Long-context multi-level	-$86.2	1.8	-0.76

Table 2: Ablation path from fill-centric quoting to direct net-outcome prediction. In the current sample, the long-context single-level model is the strongest final configuration.

At a 1.0 bps maker fee, the long-context single-level model finishes at +$21.8 over 21 days with annualized Sharpe about 1.03. Multi-level quoting remains negative even after adding context and tightening the entry threshold. That does not look like a fee issue alone. In the earlier multi-level configuration, zero-fee performance was still strongly negative, which points instead to a calibration problem under correlated depth exposure.

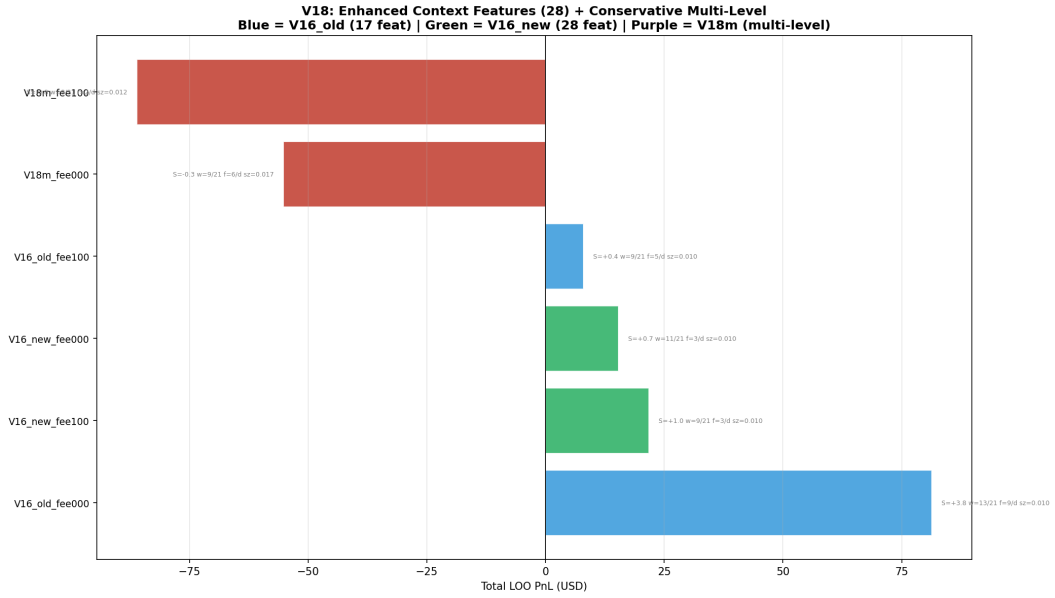

Figure 4: Comparison of short-context single-level, long-context single-level, and long-context multi-level configurations. Longer context helps the conservative single-level strategy, but it does not yet make simultaneous multi-depth quoting work.

7 Robustness and Deployment Realism

Across the different stages of the project, a few conclusions keep coming back:

1. Exponential fill structure is a good first-order empirical description for this medium-frequency setting.
2. Spread-state conditioning is not useful on BTC/USDT because the spread is almost always pinned to the tick.
3. Fill-centric quoting is economically wrong even when fill prediction itself is strong.
4. Direct economic supervision materially improves the strategy's behavior.

At the same time, the simulator is not a live-trading replica. It does not model latency competition, exact FIFO queue rank, competing market makers, or partial fills in full detail. The strategy also

benefits from a public-data backtest environment in which placement and cancellation are effectively immediate at the 100 ms observation scale. All of those omissions would tend to shrink a live edge rather than inflate it. For that reason, the +$21.8 result is better read as evidence that the objective is moving in the right direction than as a production PnL forecast.

A second realism constraint is temporal resolution. Public Binance depth updates are adequate for minute-scale execution modeling, but they are not enough for genuine HFT queue management. That limitation fits the empirical results: the model can filter out many bad orders, but it is still not fast enough to make aggressive multi-level market making robust in the most toxic states.

8 Conclusion

In this BTC/USDT medium-frequency setting, fill prediction and market-making profitability are different objectives. Fill curves are stable, fill classifiers are accurate, and the microstructure effects behave as expected. None of that is enough, on its own, to make passive market making profitable once explicit costs and inventory liquidation are included. The strategy only turns positive when the model is trained on the quantity it actually needs to rank: realized order-level net outcome.

The practical takeaway is modest but useful. At public-data speed, the best use of this work is probably not a naive standalone market-making engine. It is a selective execution layer that decides when a passive quote is worth placing at all. That is still a meaningful result: the fill model contains real signal, the economic bottleneck is genuine, and the objective function has to reflect both.

A Feature Sets

Feature set	Contents
Base fill features (10)	depth, log-depth, side, imbalance, absolute imbalance, relative spread, L1 size, log L1 size, queue-ahead, log queue-ahead
Short-context extension (17)	base features plus 10 s / 30 s / 60 s mid returns, 30 s / 60 s rolling volatility, momentum × side, depth × volatility
Long-context extension (28)	short-context features plus 5 min / 30 min returns, 5 min / 30 min volatility, micro-to-macro volatility ratio, momentum acceleration, hour-of-day sine/cosine, session volatility, running daily range, depth × 5 min volatility

Table 3: Feature families used across the empirical stages.

B Final Simulator Parameters

Parameter	Value	Role
Decision cadence	1 s	synthetic order placement frequency
Horizons	60 / 300 / 600 s	fill-label and timeout horizons
Depth grid	0.1 to 30 bps	candidate passive quote distances
Fill size	0.01 BTC	single-level order size
Inventory cap	$\pm$0.10 BTC	single-level risk limit
Inventory skew	8 bps	quote adjustment near inventory limit
Maker fee	1.0 bps	explicit cost per fill
Slippage	0.5 bps	per-fill execution penalty
Minimum tradable depth	2.0 bps	excludes quotes dominated by cost
Cooldown	10 s	wait after fill or timeout in single-level runs
Liquidation slippage	2.0 bps	end-of-day inventory close-out cost
Multi-level max slots/side	6	simultaneous live depths per side
Multi-level capital cap	5 BTC	total position constraint

Table 4: Core simulator parameters used in the final single-level and multi-level experiments.