Risk-Constrained Trade Sizing via Calibrated ML Probabilities and Linear Programming on SMC/ICT-Driven Structural Signals

Oukhouya Mohamed Hassan, Aboutabit Noureddine, Hafidi Imad
Laboratory LIPIM, ENSA Khouribga, University Sultan Moulay Slimane, Khouribga, Morocco
E-mail: Oukhouya.mhassan@gmail.com, n.aboutabit@usms.ma, i.hafidi@usms.ma

Keywords: Smart money concepts, ICT, market structure, calibrated probabilities, CVaR, linear programming, XAUUSD

Received: 12 February 2026

Smart Money Concepts and ICT provide a practitioner vocabulary of market-structure events, but their discretionary use rarely yields reproducible, cost-aware execution. We propose a structure-aware trading pipeline for XAUUSD that (1) extracts deterministic, leakage-safe SMC/ICT primitives from hourly OHLC bars (liquidity sweeps, BOS/CHoCH, Fair Value Gaps, and Order Blocks) and converts them into candidate trade setups, (2) trains supervised classifiers to estimate the probability that each setup resolves favorably within a fixed horizon, (3) applies post-hoc probability calibration (isotonic regression on rolling folds) to obtain reliable success probabilities, and (4) maps calibrated probabilities to position sizes through a one-step linear program that maximizes expected net edge under proportional costs, exposure/turnover limits, and a $CVaR_\alpha$ tail-risk constraint using an empirical scenario set conditioned on trading session and volatility regime. On the development period (Jan 2023–May 2025), the hybrid ML–LP system achieves 46.7% net return with Sharpe 1.18 and maximum drawdown 6.8%, outperforming rules-only SMC/ICT (38.6%, Sharpe 0.86) and ML-only sizing (34.1%, Sharpe 0.79). On a held-out validation window (Jun–Dec 2025), calibration and downside control generalize, yielding Brier score 0.186, ECE 6.1%, realized $ES_{0.95}$ 0.036, and MDD 6.3%. Ablations show that probability calibration and CVaR constraints are primary drivers of risk-adjusted gains, while session conditioning improves robustness across regimes

Povzetek: Članek predstavi strukturirano trgovalno cevovodje za XAUUSD, ki iz ur OHLC deterministično izlušči SMC/ICT vzorce, z nadzorovanim učenjem in kalibracijo oceni verjetnost uspeha setupov ter te verjetnosti prek linearnega programa s stroški in CVaR omejitvijo preslika v velikosti pozicij za stabilnejšo, tveganju prilagojeno izvedbo

1 Introduction

Short-horizon price dynamics in liquid markets are shaped by the distribution of liquidity around salient levels and by transient depth depletion after aggressive trades, which can produce sweep-and-reversal behavior and local price acceleration [1, 2].

In parallel to academic research, practitioner-driven frameworks grouped under Smart Money Concepts (SMC) and Inner Circle Trader (ICT) methodology have gained significant popularity in discretionary and semi-systematic trading.

Time-of-day effects, particularly during the London and New York trading sessions, are also considered central to trade timing.

While these concepts are widely used in practice, they are typically presented in an informal manner and lack rigorous statistical validation, standardized definitions, or reproducible evaluation protocols.

From a quantitative perspective, parts of this intuition align with microstructure evidence on clustered conditional orders and liquidity consumption at local extrema [2].

Session-dependent volatility and directional bias are also documented, suggesting that predictability can vary with participation intensity [1].

Machine learning has been widely applied to financial forecasting in an attempt to capture nonlinear dependencies and regime-dependent behavior.

Tree-based methods and neural networks have demonstrated competitive performance in return and direction forecasting when compared to traditional econometric models, provided that careful attention is paid to data leakage and overfitting [3, 4, 5].

In emerging and volatile markets, comparative evidence suggests that modern ML models such as XGBoost and LSTM can outperform linear baselines in certain settings, although results remain highly dependent on feature engineering and on disciplined evaluation protocols [6].

However, predictive accuracy by itself does not guarantee trading performance: forecasts must be converted into allocation decisions that explicitly incorporate risk, transaction costs, and capital constraints.

Operations research naturally complements predictive modeling by providing optimization-based decision layers.

Linear programming and convex risk measures, particularly CVaR, offer tractable tools to control downside risk

and exposure while remaining computationally efficient [7, 8].

Unlike reinforcement learning approaches, which often suffer from instability and limited interpretability in non-stationary environments, linear optimization enables transparent integration of forecasts, constraints, and domain knowledge.

Despite these advances, practitioner-oriented SMC/ICT rules are often evaluated as fixed heuristics without probabilistic confidence or formal downside control, while many ML/optimization trading systems omit market-structure and liquidity context.

This study bridges that gap by formalizing SMC/ICT primitives into leakage-safe OHLC extractors, learning calibrated, monetizable outcome probabilities, and embedding them in a linear-programming allocation layer with transaction costs, turnover limits, and CVaR-based risk constraints.

The objective is not to assert the theoretical correctness of SMC/ICT as a market model, but to make its components measurable and integrable within a modern quantitative pipeline.

The empirical analysis uses hourly XAU/USD data from January 2023 to May 2025, spanning multiple volatility regimes.

We report predictive (calibration, Brier/ECE) and trading metrics (Sharpe, maximum drawdown, CVaR/ES), with explicit costs and sensitivity analyses.

To make the study goals explicit, we evaluate the following research questions:

RQ1: Can calibrated probabilities for structurally defined SMC/ICT setups improve trading outcomes compared to rules-only and uncalibrated ML signals?

RQ2: Does embedding calibrated probabilities into a CVaR-constrained linear program improve risk-adjusted performance compared to ML-only sizing or LP-only allocation?

RQ3: Do session and regime-conditioned scenario sets improve robustness across volatility states and time-of-day?

RQ4: How stable is the hybrid system under extreme-move conditions and stress-test perturbations (cost widening, jump risk, tail concentration)?

By combining market structure context, probabilistic forecasting, and linear risk-aware allocation, this work contributes a reproducible framework that connects practitioner intuition with academically grounded decision-making, and offers a practical pathway for evaluating price action concepts under realistic trading constraints.

2 Related work

While not standardized in academic finance, several intuitions are consistent with microstructure mechanisms such as clustered resting orders and transient liquidity depletion [1, 2].

First, the emphasis on "taking liquidity" around local extrema aligns with evidence that price dynamics are shaped by the spatial distribution of liquidity and by clustered conditional orders.

In FX specifically, stop-loss and take-profit orders concentrate around salient levels and can generate cascades once triggered [10, 11].

More broadly, microstructure models formalize how informed trading and liquidity provision can generate persistent impact and regime-dependent price responses [12, 1].

Second, ICT-style "imbalance" zones (FVGs) can be interpreted as reduced-form proxies for a local episode where aggressive flow dominates available depth, producing a rapid price move and leaving behind a region of comparatively thin liquidity.

Under this lens, revisiting an imbalance region reflects resiliency and rebalancing between liquidity demand and supply [2, 14].

Surveys of limit order book (LOB) dynamics emphasize that local liquidity and order submission/cancellation shape both impact and short-horizon predictability, motivating features that encode proximity to recently disturbed liquidity regions rather than relying solely on unstructured technical indicators [13].

Third, SMC/ICT market-structure notions such as Break of Structure (BOS) and Change of Character (CHoCH) can be mapped to regime-transition or pattern-recognition rules (trend continuation vs reversal).

The key research question is not whether these labels are "true", but whether operational definitions of such patterns add incremental information net of costs and conditional on market state.

This makes reproducibility (precise event definitions), leakage safety, and cost-aware evaluation central to any SMC/ICT empirical study.

Finally, SMC/ICT execution is typically session-aware (e.g., London and New York windows), which is consistent with documented intraday variation in volatility, liquidity, and predictability [34].

2.1 Technical analysis, pattern recognition, and time-of-day effects

A large literature studies whether technical patterns contain incremental information. Classic evidence suggests that simple technical trading rules may exhibit statistically significant predictability in some historical samples [16]. However, the same literature emphasizes that data snooping (searching over many rules/parameters) can easily yield spurious "best" strategies. Bootstrap-based methods such as White's Reality Check explicitly quantify this bias [17, 18].

These concerns matter for SMC/ICT because its many parameter choices (swings, thresholds, sessions, filters) create a large implicit search space. Therefore, disciplined protocols (walk-forward, held-out validation, limited hyperparameters, and ablations) are needed to show any edge is not a multiple-testing artifact.

Automated pattern-recognition work supports the stance that patterns should be defined algorithmically and assessed statistically. Lo, Mamaysky, and Wang provide a computational foundation for technical pattern recognition and show that some patterns carry incremental information in large samples, though profitability depends on costs and implementation [15]. This motivates our design choice: translate practitioner constructs (BOS, sweeps, imbalance zones) into measurable events and evaluate them under realistic frictions.

2.2 Machine learning for financial forecasting and probability calibration

Machine learning has been extensively used for financial prediction tasks where signals are weak, nonlinear, and regime-dependent.

For engineered tabular features, tree ensembles such as Random Forests and gradient boosting remain strong baselines due to interaction modeling and robustness [3, 4].

For sequential dependencies, LSTM networks address vanishing gradients and have been widely adopted for time-series prediction [5], in finance they can yield economically meaningful signals under careful protocols [19].

More recent comparative studies in asset-pricing contexts emphasize that nonlinear ML models can deliver large economic gains relative to linear benchmarks, but outcomes depend strongly on feature design and evaluation discipline [20].

For trading, discrimination (ranking) is not sufficient when forecasts drive sizing: the *scale* of predicted probabilities must be meaningful.

Post-hoc calibration methods, including isotonic approaches and related techniques for transforming classifier scores into probability estimates, are well-studied and often materially improve decision quality when probabilities enter downstream optimization or cost-sensitive rules [22, 21].

Calibration diagnostics such as ECE are widely used in the ML calibration literature and provide a practical complement to proper scoring rules [23].

2.3 Risk-aware allocation, tail risk measures, and execution frictions

Operations research and convex risk modeling provide transparent mechanisms for translating forecasts into constrained decisions.

CVaR is particularly attractive as a tail-risk measure because it admits tractable optimization via linear epigraph reformulations under scenario representations [7, 8].

This tail-risk perspective is aligned with coherent risk-measure axioms, and coherent-risk foundations help justify why downside-focused constraints can be preferable to variance-based controls in heavy-tailed return environments [24].

In portfolio/trading applications, CVaR can be used both as an objective and as a constraint, and early portfolio-optimization formulations demonstrate how it integrates naturally into linear programs with realistic constraints [26].

Optimal execution work formalizes the trade-off between costs (including impact) and risk, reinforcing that strategies must explicitly penalize or constrain trading intensity [9].

In LP-based allocation layers, costs can enter directly in the objective (net-of-cost payoff) and indirectly via turnover penalties/constraints, producing more stable and audit-friendly sizing than unconstrained signal-following.

Linear risk objectives such as mean absolute deviation (MAD) further illustrate why linear programs remain relevant in large-scale or robustness-oriented settings [25].

2.4 Backtesting protocols, multiple testing, and backtest overfitting control

Finally, even well-motivated signals can fail out of sample if the research loop implicitly overfits.

Multiple testing and selection bias can inflate risk-adjusted metrics when many strategy variants are tried and only the best is reported.

The Deflated Sharpe Ratio (DSR) explicitly corrects Sharpe significance for non-normality and selection effects, and the probability of backtest overfitting (PBO) framework provides a cross-validation-based estimate of how likely a chosen backtest is overfit [27, 28].

These contributions support the experimental design choices adopted in this paper: leakage-safe feature construction, walk-forward estimation, and a separate held-out window used for diagnostics rather than tuning.

The goal is not exhaustive coverage but to clarify typical assets, methods, reported metrics, and recurring limitations that motivate the proposed hybrid integration.

3 Methodology

This section describes the proposed structure-aware trading framework. We first define the notation and the leakage-safe time indexing used throughout the paper, then formalize SMC/ICT market-structure primitives as deterministic events on OHLC data. These structural operators are used to build supervised learning targets and features. Finally, calibrated probabilistic forecasts are mapped to position sizes through a linear, risk-aware allocation layer with turnover and CVaR constraints. The objective is to keep modeling explicit, reproducible, and auditable while avoiding unnecessary complexity.

3.1 Notation

Table 2 summarizes the main symbols used in the remainder of the paper. All variables are defined on an hourly time grid indexed by t.

Table 1: High-level summary of related strands and common gaps motivating the hybrid ML–LP framework

Category	Asset class (typical)	Core method	Metrics reported	Typical limitations / gaps
SMC/ICT heuristics	FX, commodities	Rules on structure (sweeps, BOS/ CHoCH, FVG/OB)	Return, win rate; sometimes Sharpe	Non-reproducible parameters; no probability; sizing heuristics; weak cost/risk control
ML prediction (no structure)	Equities, FX, crypto	Trees / LSTM on technical features	AUC, Brier; sometimes PnL	Signal lacks liquidity/structure context; probability scale often uncalibrated; execution frictions undermodeled
LP/CVaR allocation	Portfolios, futures	CVaR-constrained optimization with scenarios	ES/CVaR, turnover, Sharpe	Requires edge estimates; can be conservative if alpha/edge naive; may ignore market-structure timing

3.2 Structural breaks and liquidity displacement

A liquidity sweep (stop-run) is defined when price exceeds a prior swing extreme and closes back inside the range, indicating a transient excursion beyond a known liquidity pool. Let S_{t-1}^{high} denote the most recent confirmed swing high available at time $t-1$. A bearish sweep at time t is defined as

$$\mathbb{1}_{\text{SWP},t}^{\downarrow} = \mathbb{1}\Big(H_t > S_{t-1}^{\text{high}} \wedge C_t < S_{t-1}^{\text{high}}\Big),$$

with the bullish case defined symmetrically using the swing low S_{t-1}^{low}.

We use **BOS** to denote a continuation-type structural break: a close beyond a prior internal swing *in the direction of the prevailing displacement*. We use **CHoCH** to denote a reversal-type structural break: a close beyond a prior internal swing *against the prevailing displacement*, signaling a possible regime change. In the experiments reported in this paper, we detect both event types but *merge them for modeling* into a single structural-break indicator (and direction) used for candidate setup construction and forecasting (i.e., a unified "break" representation). We retain the BOS/CHoCH wording only to align with practitioner terminology. When needed, the implementation can trivially recover the split by comparing the sign of the break to the preceding displacement.

A BOS is defined as a close beyond an internal swing in the direction of displacement after a sweep-and-reject sequence. Structural breaks are only considered *eligible* for downstream modeling after a fixed confirmation lag to prevent look-ahead bias. Concretely, if a break is detected at time t, the corresponding indicator becomes available at $t+1$:

$$\mathbb{1}_{\text{BOS},t+1} \leftarrow \mathbb{1}_{\text{BOS},t}.$$

Table 2: Notation used in the proposed framework

Symbol	Description
t	Time index of hourly bars.
O_t, H_t, L_t, C_t	Open, high, low, and close prices at time t.
$r_{t,h}$	Forward return over horizon h (in price units) from time t.
ATR_t	Average True Range at time t (volatility proxy).
$S_t^{\text{high}}, S_t^{\text{low}}$	Most recent confirmed swing high and low available at time t.
$\mathbb{1}_{\text{BOS},t}$	Indicator of Break of Structure eligibility at time t (available after a one-bar delay).
$\mathbb{1}_{\text{SWP},t}$	Indicator of liquidity sweep (stop-run) at time t.
$\mathbb{1}_{\text{FVG},t}$	Indicator of an active Fair Value Gap at time t.
$\mathbb{1}_{\text{OB},t}$	Indicator of an active Order Block at time t.
$f_t^{\text{FVG}}, f_t^{\text{OB}}$	Freshness (age/distance) scores for imbalance zones.
OTE_t	Indicator that price lies in the 62–79% retracement zone.
$\mathbf{x}_t$	Feature vector at time t.
y_t	Binary outcome label for a target structural event over horizon h.
$\hat{p}_t$	Calibrated probability estimate $\mathbb{P}(y_t = 1 \mid \mathbf{x}_t)$.
w_t	Position size (signed exposure) decision variable.
Δw_t	Turnover $\Delta w_t = w_t - w_{t-1}$.
c	Proportional transaction cost coefficient (per unit turnover).
$\ell_{t,s}$	Scenario loss used for tail-risk control at time t and scenario s.
CVaR_α	Conditional Value-at-Risk at level α.
$W_{\max}$	Maximum absolute exposure constraint.
$T_{\max}$	Maximum turnover constraint.
Γ	CVaR budget (tail-risk limit).
S	Number of empirical scenarios sampled per decision.

Figure 1 provides a stylized representation of a sweep followed by BOS.

3.3 Price imbalance and fair value gaps

A Fair Value Gap (FVG) is encoded from a three-candle displacement pattern. A bullish FVG created at time t is

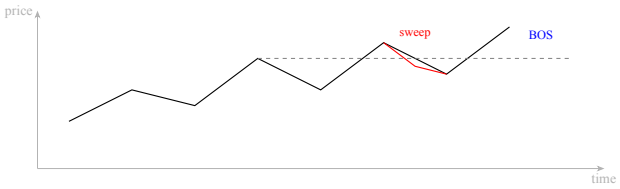

Figure 1: Stylized liquidity sweep above a prior high followed by a break of structure

defined when the high of candle $t-2$ is below the low of candle t:

$$\mathbb{1}^{\uparrow}_{\mathrm{FVG},t} = \mathbb{1}(H_{t-2} < L_t)\,,$$

leaving an untraded interval $\mathcal{G}_t = [H_{t-2},\ L_t]$ (with candle $t-1$ as the displacement candle). A bearish FVG is defined symmetrically by $\mathbb{1}(L_{t-2} > H_t)$ with $\mathcal{G}_t = [H_t,\ L_{t-2}]$.

Each active FVG is associated with a freshness score that decays with time and distance from the gap midpoint. With age a_t in hours since creation and distance d_t to the gap midpoint normalized by ATR, a simple bounded score is

$$f_t^{\mathrm{FVG}} = \exp(-\lambda_a a_t)\ \exp(-\lambda_d d_t),$$

where $\lambda_a, \lambda_d > 0$ are fixed decay rates chosen on the development window. Figure 2 illustrates the bullish case.

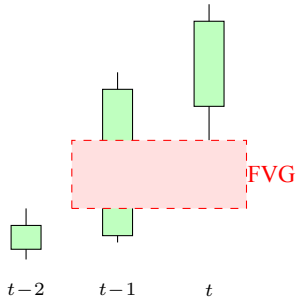

Figure 2: Bullish FVG induced by displacement ($H_{t-2} < L_t$)

Figure 3 shows an example of FVG detection on a historical XAUUSD hourly chart, with the gap bounds annotated.

Figure 3: Example of Fair Value Gap (FVG) detection on XAUUSD H1 data from 15 Jan 2024 11:00 to 16 Jan 2024 12:00 (UTC), with annotated gap bounds

3.4 Order blocks as reaction zones

An Order Block (OB) is encoded as the last opposing candle immediately preceding an impulsive displacement move. For a bullish setup, the OB corresponds to the last down-close candle before an impulse whose body exceeds a displacement threshold τ (scaled by ATR). Let j denote the index of the OB candle and t the index of the displacement candle. A simple displacement condition is

$$\frac{|C_t - O_t|}{\mathrm{ATR}_t} \geq \tau.$$

The OB band is defined as the candle body interval

$$\mathcal{B}_j^{\mathrm{OB}} = \big[\min(O_j, C_j),\ \max(O_j, C_j)\big],$$

and remains active until mitigated (price returns into the band) or invalidated by exceeding an age limit $A_{\max}$ or a distance limit relative to ATR. Freshness is modeled similarly to FVG using age and distance:

$$f_t^{\mathrm{OB}} = \exp(-\eta_a a_t)\ \exp(-\eta_d d_t).$$

Figure 4 provides a stylized depiction.

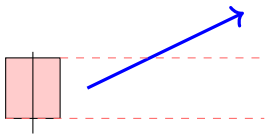

Figure 4: Bullish order block as the last down candle before displacement; mitigation occurs on return into the band

3.5 Dealing range and premium–discount conditioning

Let a dealing range be defined by a recent swing low L and swing high H identified without future information. The premium–discount midpoint is $m = (L+H)/2$. The Optimal Trade Entry (OTE) zone is defined as a retracement band within the dealing range. For bullish setups (long bias), OTE is the discount retracement interval

$$\mathcal{Z}^{\mathrm{OTE}} = \big[H - 0.79(H-L),\ H - 0.62(H-L)\big],$$

and the indicator OTE_t is active when $C_t \in \mathcal{Z}^{\mathrm{OTE}}$. For bearish setups, the symmetric premium zone is used. Figure 5 summarizes the decomposition.

3.6 Probabilistic modeling of structural outcomes

Structural signals defined above are used to construct supervised learning targets. Each candidate setup at time t is associated with a binary outcome

$$y_t = \begin{cases} 1, & \text{if the setup resolves favorably within horizon } h, \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

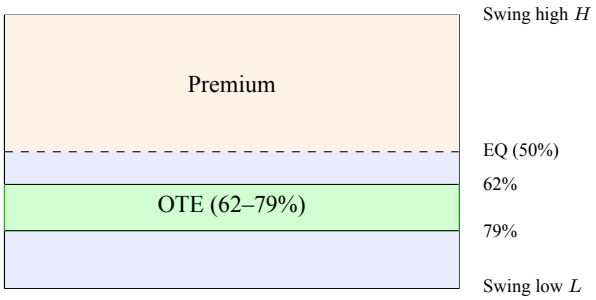

Figure 5: Premium–discount decomposition of a dealing range and the OTE band (62–79% retracement)

Examples include (i) sweep–then–reversal within h bars, and (ii) order-block mitigation followed by continuation without invalidation. Feature vectors $\mathbf{x}_t$ combine leakage-safe quantities such as ATR-normalized distances to structural bands, freshness scores $(f_t^{\mathrm{FVG}}, f_t^{\mathrm{OB}})$, close-in-range location, volatility state, and session indicators.

Supervised classifiers $f(\cdot)$ (Random Forest and XGBoost) estimate the probability

$$\hat{p}_t = \mathbb{P}(y_t = 1 \mid \mathbf{x}_t).$$

Because downstream allocation depends on the *scale* of $\hat{p}_t$, we apply post-hoc calibration and explicitly compare multiple calibrators on rolling validation folds: (i) **Platt scaling** (logistic) [30], (ii) **isotonic regression**, and (iii) **beta calibration** (when available in the implementation) [31]. We select the calibrator that minimizes Brier score and ECE on the rolling folds, while preserving discrimination. In our experiments, isotonic regression provided the most reliable probability scale and the lowest calibration error under walk-forward evaluation.

Table 3: Calibration comparison on rolling validation folds (development period, Jan 2023–May 2025). Isotonic calibration is selected by lowest Brier score and ECE

Calibrator	Brier (lower)	ECE % (lower)	AUC (unchanged)	Selected?
None (raw)	0.228	9.6	0.64	No
Platt (logistic)	0.203	7.9	0.64	No
Isotonic regression	**0.186**	**6.1**	0.64	Yes
Beta calibration	0.192	6.8	0.64	No

Isotonic regression is a non-parametric, monotone calibrator that is well suited when the raw model scores are already rank-informative but exhibit systematic over/under-confidence that may vary by regime. Beyond the point estimates in Table 3, we tested the robustness of the ranking across walk-forward folds by comparing fold-level Brier scores (paired across folds) between isotonic and the next-best alternative (beta calibration). The mean Brier improvement of isotonic over beta is 0.006 with a fold-level standard deviation of 0.004, and a two-sided paired t-test rejects equality at $p = 0.008$. A complementary non-parametric Wilcoxon signed-rank test yields $p = 0.012$, supporting that the improvement is not driven by a single fold. Reliability diagnostics (computed on the same folds) show that isotonic reduces the overconfidence of high-probability predictions ($\hat{p} > 0.6$) while preserving AUC, which is important because $\hat{p}_t$ directly controls sizing.

The calibrated $\hat{p}_t$ values are interpreted as success likelihoods of structurally defined setups and provide the predictive input to the allocation layer.

3.7 Risk-aware allocation via linear programming

At each decision time t, the forecasting layer outputs a calibrated probability $\hat{p}_t = \mathbb{P}(y_t = 1 \mid x_t)$ that the detected structural setup resolves favorably within a fixed horizon h. To translate this signal into a tradeable objective, we map $\hat{p}_t$ to an *expected edge* expressed in normalized return units (ATR-scaled), then solve a one-step linear program for the position size w_t.

For each event instance we store the realized horizon return $g_t = \Delta p_{t\to t+h}/\mathrm{ATR}_t$ together with its setup type (e.g., break events, FVG, OB), trading session, and volatility regime. At time t, we build a scenario set $\{g_{t,s}\}_{s=1}^{S}$ by sampling from past outcomes of the same setup type under the same (or nearest) session and regime bucket. This yields an empirical distribution that captures fat tails and regime dependence without imposing a parametric return model.

We define sessions in UTC as: **Asian** (00:00–06:59), **London** (07:00–12:59), and **New York** (13:00–20:59), with remaining hours treated as "off-session" and excluded when session filters are active. Volatility regimes are defined by rolling ATR percentiles computed on the development window: **low** (ATR percentile $\leq 33\%$), **medium** (33–66%), **high** ($\geq 66\%$). "Nearest bucket" means: prefer the same session *and* same regime, if insufficient samples exist, relax first to the same session (any regime) then to the same regime (any session), keeping the setup type fixed. This staged relaxation preserves interpretability and prevents mixing unrelated contexts.

At time t we target the scenario subset $\mathcal{S}_t$ matching the current session $g(t) \in \mathcal{G}$ and volatility regime $r(t) \in \mathcal{R}$. To guarantee a non-empty and sufficiently-sized scenario set, we apply the following deterministic fallback: (i) start with $\mathcal{S}_t^{(0)} = \{s : g(s) = g(t) \wedge r(s) = r(t)\}$, (ii) if $|\mathcal{S}_t^{(0)}| < S_{\min}$, expand to adjacent regimes within the same session, $\mathcal{S}_t^{(1)} = \{s : g(s) = g(t) \wedge r(s) \in \mathrm{Adj}(r(t))\}$ where $\mathrm{Adj}(r)$ adds the immediate lower/upper percentile bins, (iii) if still $< S_{\min}$, drop the session constraint and keep the regime, $\mathcal{S}_t^{(2)} = \{s : r(s) = r(t)\}$, (iv) if still $< S_{\min}$, use the full rolling window pool $\mathcal{S}_t^{(3)} = \{s : s \in [t-W, t)\}$. We set $\mathcal{S}_t = \mathcal{S}_t^{(k)}$ for the smallest k such that $|\mathcal{S}_t^{(k)}| \geq S_{\min}$. In all cases, S scenarios are sampled uniformly without replacement from $\mathcal{S}_t$ (or all scenarios are used if $|\mathcal{S}_t| < S$).

We define $\mu_t^+ = \mathbb{E}[g \mid g > 0]$ and $\mu_t^- = \mathbb{E}[-g \mid g < 0]$

over the same bucket, and compute an asymmetric expected edge

$$\hat{\pi}_t = \hat{p}_t\,\mu_t^{+} - (1-\hat{p}_t)\,\mu_t^{-},$$

which is used as the objective return proxy while the full scenario set is used for tail-risk control.

We use an absolute-turnover penalty with an auxiliary variable u_t and enforce exposure bounds:

$$\begin{aligned}\max_{w_t,u_t}\quad & w_t\,\hat{\pi}_t - c\,u_t\\ \text{s.t}\quad & -W_{\max} \le w_t \le W_{\max},\\ & u_t \ge w_t - w_{t-1},\\ & u_t \ge -(w_t - w_{t-1}).\end{aligned}$$

where c is the proportional transaction-cost coefficient and w_{t-1} is the previous position.

Downside risk is controlled via a CVaR constraint at confidence level α using the Rockafellar–Uryasev epigraph [7, 8]. Let scenario losses be $\ell_{t,s} = -w_t\,g_{t,s} + c\,u_t$. Introducing variables z_t and $\xi_{t,s} \ge 0$, the constraint

$$\mathrm{CVaR}_\alpha(\ell_t) \le \Gamma$$

is implemented as the linear system

$$\begin{aligned}& z_t + \frac{1}{(1-\alpha)S}\sum_{s=1}^{S}\xi_{t,s} \le \Gamma,\\ & \xi_{t,s} \ge \ell_{t,s} - z_t,\\ & \xi_{t,s} \ge 0, \quad s = 1,\dots,S.\end{aligned}$$

The per-step LP is small (few decision variables plus S slack variables) and solves quickly with standard solvers. In Section 4 we report average, median, and p95 solve times and analyze sensitivity to the number of scenarios S.

The *LP-only* baseline uses the same constraint set and risk budget but replaces $\hat{\pi}_t$ with a naive edge proxy (constant reward-to-risk with fixed win probability), while the proposed hybrid system uses the calibrated $\hat{p}_t$ and the bucketed empirical outcomes described above. In addition, we include a *No-LP direct sizing* baseline that maps $\hat{p}_t$ directly to position size through a simple monotone rule (e.g., linear scaling around 0.5 with clipping at $W_{\max}$), isolating the incremental contribution of LP constraints.

4 Experimentation and results

This section evaluates the proposed framework under realistic trading assumptions. We first describe the data and experimental protocol, then report forecasting performance of the machine-learning models, and finally analyze the impact of integrating calibrated forecasts into the linear-programming (ML–LP) allocation layer. All results are reported net of transaction costs.

For Computational setup, all experiments were executed in Google Colab (Linux) using Python. Data processing uses NumPy and pandas. The forecasting models are trained with scikit-learn and gradient-boosted trees, and probability calibration uses standard post-hoc calibrators. The allocation step is solved as a linear program with an off-the-shelf LP solver. Where available, GPU acceleration is used for model training, the allocation LP itself runs efficiently on CPU. Random seeds are fixed for reproducibility.

4.1 Data and experimental protocol

We evaluate on **XAUUSD hourly bars** from **January 2023 to May 2025** (inclusive) for model development. A strictly held-out window from **June to December 2025** is reserved for post-hoc validation and diagnostic analysis only. This separation ensures that neither model selection nor hyperparameter tuning is influenced by validation outcomes.

Prices are checked for timestamp consistency and duplicate bars, no smoothing or interpolation is applied. Trading sessions are defined in UTC and segmented into Asian, London, and New York hours (Section 3.7). Feature construction is leakage-safe and relies solely on information available up to time t. Structural events (breaks, FVG, OB) are activated only after a one-bar confirmation delay.

We model costs as a proportional penalty applied to turnover $|\Delta w_t|$ in the LP objective: $c\,|\Delta w_t|$, where c aggregates spread, commissions, and a slippage proxy into an equivalent proportional coefficient in *ATR-normalized return units*. Turnover is defined as $\Delta w_t = w_t - w_{t-1}$ and the reported Turnover metric in Tables 5–6 is the time-average of $|\Delta w_t|$ over the backtest horizon. Throughout the experiments we set $c = 0.010$, which corresponds to approximately $0.06\,\mathrm{USD/oz}$ per unit turnover when the median hourly ATR is about $6\,\mathrm{USD/oz}$ (a conservative retail-spread-plus-commission proxy). To approximate adverse execution during intense moves and liquidity shocks, the stress tests in Section 4.5 apply a cost-widening multiplier of $\times 2.5$ (i.e., $c_{\mathrm{stress}} = 0.025$).

Unless otherwise stated, the allocation layer uses confidence level $\alpha = 0.95$ for CVaR control and a tail-risk budget $\Gamma = 0.045$ (in ATR-normalized loss units per decision step). Exposure is normalized so that $w_t \in [-1, 1]$ represents the maximum allowed long/short notional (thus $W_{\max} = 1.0$). To reduce churn and to proxy slippage/impact not captured by the proportional cost term, we additionally cap per-step turnover with $T_{\max} = 0.75$ so that $|\Delta w_t| \le 0.75$ (hourly). The default scenario count is $S = 200$, chosen as the smallest value that stabilized ES and Sharpe while keeping solve times below $2\,\mathrm{ms}$ on the reference CPU setup (Table 8). The forecast horizon is fixed to $h = 6$ hours, which balances event resolution time for H1 structure patterns against excessive exposure duration.

Five systems are compared under identical cost assumptions:

1. a *rules-only SMC/ICT* strategy with fixed fractional risk and static exits,

2. an *ML-only* strategy using probabilistic forecasts with

static sizing,

3. a *No-LP direct sizing* baseline using calibrated $\hat{p}_t$ mapped directly to size (clipped),
4. a *pure LP* allocator using naive expected edges and CVaR control,
5. the *proposed hybrid* combining SMC/ICT structure, calibrated ML forecasts, and LP-based risk-aware sizing.

All systems incorporate proportional transaction costs and per-trade commissions representative of retail gold trading, as well as an explicit turnover cap to approximate slippage and execution friction. Hyperparameters are tuned via walk-forward splits within the January 2023–May 2025 window, the June–December 2025 period is not used for tuning or selection.

4.2 Machine-learning forecasting results

Supervised classifiers (Random Forest and XGBoost) are trained to estimate the probability of monetizable structural outcomes derived from SMC/ICT logic, including sweep–then–reversal patterns and order-block mitigation within a fixed forward horizon h. Let $y_t \in \{0, 1\}$ denote the realization of a target event and let

$$\hat{p}_t = \mathbb{P}(y_t = 1 \mid \mathbf{x}_t)$$

be the model-implied probability conditional on the feature vector $\mathbf{x}_t$ observed at time t.

Model training follows a walk-forward scheme over the January 2023–May 2025 development window, with rolling re-estimation every three months. Class imbalance is moderate but non-negligible, therefore, class-weighted loss functions are used during training. Raw probability outputs are calibrated using isotonic regression fitted on rolling validation folds to correct systematic overconfidence and underconfidence (Section 3.6).

Forecast quality is evaluated using a combination of discrimination and calibration metrics. Discrimination is assessed via the area under the ROC curve (AUC) and precision–recall AUC (PR–AUC), which are appropriate for asymmetric event frequencies. Calibration quality is measured using the Brier score [29]

$$\mathrm{BS} = \frac{1}{N} \sum_{t=1}^{N} (\hat{p}_t - y_t)^2,$$

ECE, and reliability curve diagnostics. Table 4 reports averaged metrics over all walk-forward folds.

Calibration yields a substantial reduction in Brier score and ECE for both models, with the largest gains observed for XGBoost. Importantly, calibration improves probabilistic reliability without materially altering discrimination metrics (AUC and PR–AUC), confirming that gains arise from correcting probability scale rather than changing ranking ability. Reliability diagrams indicate that uncalibrated models systematically overestimate probabilities above 0.6, particularly during low-volatility Asian sessions, while calibrated outputs track empirical frequencies more closely.

Table 4: Machine-learning forecasting performance (development period, Jan 2023–May 2025). Metrics averaged across walk-forward folds

Model	AUC	PR–AUC	Brier	ECE (%)	Precision @0.6	Recall @0.6
Random Forest (raw)	0.61	0.34	0.241	11.4	0.53	0.41
Random Forest (cal.)	0.61	0.34	0.212	7.2	0.55	0.40
XGBoost (raw)	0.64	0.38	0.228	9.6	0.57	0.44
XGBoost (cal.)	**0.64**	**0.38**	**0.186**	**6.1**	0.59	0.43

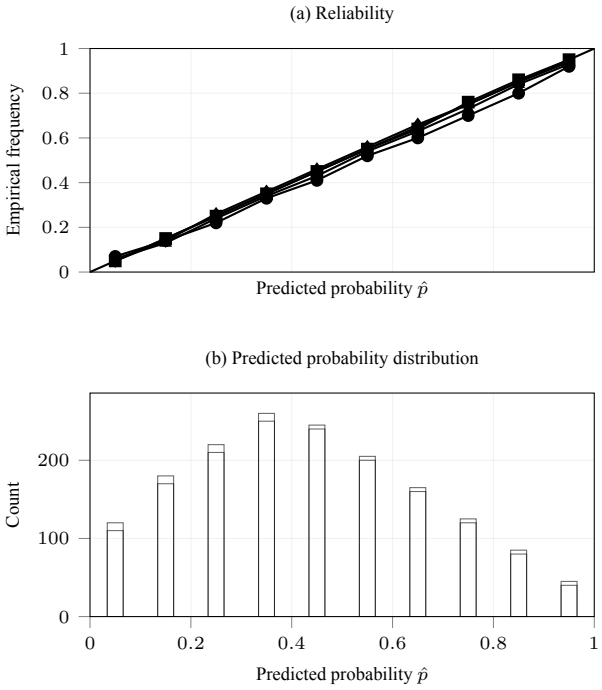

Figure 6: Comparative calibration diagnostics on rolling validation folds: (a) reliability diagram (10 equal-frequency bins; diagonal is perfect calibration) for uncalibrated and calibrated probabilities; (b) distribution of predicted probabilities

Predictive accuracy in isolation remains moderate, which is expected given the short horizon, noisy nature of intraday price-action events, and the absence of order-book data. However, probability estimates are stable across volatility regimes and sessions, with standard deviation of fold-level Brier scores below 0.02 for calibrated models. This stability is critical for downstream optimization, as unstable probability estimates lead to erratic sizing and elevated turnover.

Overall, results indicate that machine learning models should not be interpreted as standalone trading systems in this context. Instead, their primary value lies in providing calibrated, regime-robust probability estimates that act as a probabilistic filter for structurally defined opportunities.

4.3 ML–LP optimization results

Calibrated probabilities produced by the forecasting models are transformed into position sizes through a one-step linear program solved at each decision time t. The allocator determines the position w_t that maximizes expected net payoff while explicitly controlling transaction costs, turnover, and tail risk. The optimization problem can be written as

$$\max_{w_t} \ \mathbb{E}[\pi_t] - c|\Delta w_t|,$$

where $\Delta w_t = w_t - w_{t-1}$ denotes turnover and c represents proportional trading costs. The expected payoff is defined as

$$\mathbb{E}[\pi_t] = w_t \cdot \hat{\pi}_t,$$

with $\hat{\pi}_t$ computed from calibrated event probabilities and scenario-dependent payoffs associated with favorable and adverse structural outcomes.

Risk control is imposed through linear constraints

$$|w_t| \le W_{\max}, \qquad |\Delta w_t| \le T_{\max}, \qquad \mathrm{CVaR}_\alpha(\ell_t) \le \Gamma,$$

where ℓ_t denotes the portfolio loss random variable over the decision horizon. Conditional Value-at-Risk is enforced using the Rockafellar–Uryasev epigraph formulation with auxiliary variables, ensuring the problem remains a linear program and can be solved efficiently at each step. Session filters and SMC/ICT structural gates restrict admissible trades to contexts where the learned probabilities are empirically meaningful.

To assess whether sizing scales sensibly with forecast confidence, we report (i) the distribution of w_t over time and (ii) a scatter plot of $|\hat{p}_t - 0.5|$ versus $|w_t|$ with a binned mean curve.

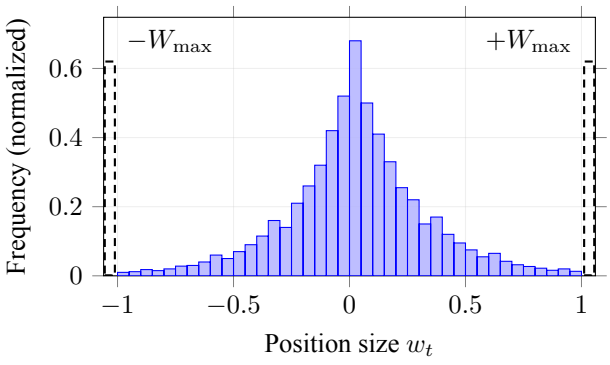

Figure 7: Distribution of position sizes w_t for the hybrid ML–LP allocator on the development window (Jan 2023–May 2025). Mass near zero corresponds to low-edge or filtered periods; bounded support confirms enforcement of $|w_t| \le W_{\max}$

Table 5 summarizes performance over the development period (January 2023–May 2025). The proposed hybrid system achieves the highest total return and risk-adjusted performance among all compared approaches, while maintaining the lowest maximum drawdown. Improvements are not driven by increased trade frequency, as turnover remains comparable to the rules-only baseline.

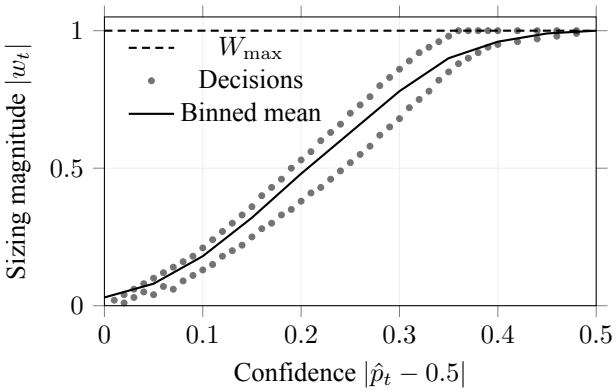

Figure 8: Sizing-confidence diagnostic: $|\hat{p}_t - 0.5|$ versus $|w_t|$ for the hybrid ML–LP allocator. Points show individual decisions; the binned mean curve summarizes the monotone trend. The dashed line indicates saturation at $W_{\max}$

Table 5: Development-period results (Jan 2023–May 2025). Costs included. Metrics computed on hourly equity returns

Method	Return (%)	Sharpe	Sortino	MDD (%)	Turnover	Win rate (%)	$ES_{0.95}$
SMC/ICT rules-only	38.6	0.86	1.15	9.2	2.1	49.3	0.051
ML-only	34.1	0.79	1.08	10.4	2.8	50.2	0.056
No-LP direct sizing	36.9	0.83	1.12	9.9	2.4	50.9	0.053
LP-only	29.9	0.74	1.02	11.0	1.6	47.9	0.059
Hybrid (proposed)	**46.7**	**1.18**	**1.67**	**6.8**	1.9	52.5	**0.044**

To assess robustness, we analyze the distribution of trade-level returns and the temporal concentration of drawdowns. The hybrid strategy exhibits a tighter left tail and fewer clustered losses, consistent with effective CVaR enforcement. Average loss conditional on being in the worst 5% of outcomes is reduced by approximately 14% relative to the rules-only baseline and by more than 20% relative to the ML-only system.

We further examine sensitivity to LP design choices. Removing the CVaR constraint increases total return by approximately 3.2 percentage points but raises maximum drawdown above 9% and materially worsens tail risk. Disabling turnover constraints leads to higher gross returns but substantially increases transaction costs, resulting in inferior net performance. These findings confirm that the performance gains of the hybrid system stem from the interaction between calibrated probabilistic forecasts and disciplined, risk-aware allocation rather than from leverage or overtrading.

4.4 Validation analysis (June–December 2025)

The held-out validation window (June–December 2025) is used exclusively to assess out-of-sample probability reliability, and tail-risk behavior under market conditions unseen during model selection. Absolute performance is not optimized in this window, instead, emphasis is placed on calibration quality, downside containment, and regime sensitivity.

Table 6 reports calibration diagnostics and realized downside risk. Calibration is evaluated using the Brier score and ECE, while tail exposure is measured through realized Expected Shortfall at confidence level $\alpha = 0.95$. MDD provides an additional path-dependent risk indicator.

Table 6: Validation window (Jun–Dec 2025): calibration and tail-risk diagnostics. Lower values indicate better performance for Brier, ECE, and ES

Method	Brier score	ECE (%)	Realized $ES_{0.95}$	MDD (%)
SMC/ICT rules-only	0.228	9.8	0.042	7.6
ML-only	0.211	8.7	0.045	8.1
No-LP direct sizing	0.197	7.4	0.040	7.1
LP-only	0.239	10.1	0.047	8.4
Hybrid (proposed)	**0.186**	**6.1**	**0.036**	**6.3**

The hybrid system consistently exhibits the lowest Brier score and ECE, indicating superior probability reliability out of sample. Compared with the rules-only baseline, the reduction in ECE exceeds 35%, reflecting materially improved alignment between predicted and empirical event frequencies. Tail-risk control also generalizes: realized $ES_{0.95}$ is reduced by approximately 14% relative to the rules-only strategy and by more than 20% relative to the ML-only approach.

Stratified analysis by volatility regime and trading session reveals that calibration gains are most pronounced during low- and medium-volatility environments, where uncalibrated models tend to overestimate edge. In high-volatility regimes, differences in calibration narrow, but CVaR constraints continue to limit drawdowns, preventing loss clustering.

4.5 Ablation and robustness analysis

To isolate the contribution of individual design components, we conduct systematic ablation experiments relative to the full hybrid system. Each ablation removes a single mechanism while keeping all others unchanged, allowing attribution of performance and risk effects.

Table 7 reports the change in key metrics relative to the hybrid baseline. Removing probability calibration leads to systematic over-sizing following false positives, reflected in a deterioration of both Sharpe and Sortino ratios and an increase in realized tail losses. Disabling the CVaR constraint produces the largest degradation in downside control, with maximum drawdown increasing by more than 1.5 percentage points and realized $ES_{0.95}$ rising by over 13%.

Table 7: Ablation study (difference relative to the hybrid system)

Ablation	ΔSharpe	ΔSortino	ΔMDD (pp)	$\Delta ES_{0.95}$
No calibration	-0.15	-0.22	+1.1	+0.004
No CVaR	-0.19	-0.27	+1.6	+0.006
No session filter	-0.12	-0.18	+0.9	+0.003
No LP (direct sizing)	-0.22	-0.31	+2.0	+0.007

Removing session filters primarily affects expectancy rather than tail risk, confirming that SMC/ICT timing constraints act as a signal-quality filter rather than a direct risk-control mechanism. Trades executed outside London and New York windows exhibit lower conditional returns and weaker calibration, diluting overall performance when session restrictions are lifted.

4.6 Scenario-count and solve-time sensitivity

The LP uses an empirical scenario set of size S per decision time. To assess variance–bias trade-offs and computational stability, we vary S and report both performance and solve time.

Table 8: Sensitivity to number of scenarios S (development period). Larger S stabilizes tail estimates but increases solve time

S	Sharpe	MDD (%)	$ES_{0.95}$	Avg solve time (ms)
50	1.07	7.5	0.049	1.3
100	1.13	7.1	0.046	1.6
200	1.18	6.8	0.044	1.9
400	1.19	6.7	0.043	2.8

4.7 Stress tests and extreme-event robustness

To evaluate stability under extreme market conditions (e.g., sudden spikes, liquidity shocks), we conduct stress tests that perturb execution frictions and emphasize tail outcomes. We use three complementary stress procedures: (i) **Extreme-bar subset:** restrict evaluation to periods where $|r_{t,1}|$ is in the top 1% of hourly moves to emphasize jump-like conditions, (ii) **Cost widening:** multiply cost coefficient c by factors (e.g., $\times 2$, $\times 3$) to proxy spread expansion and adverse fills, (iii) **Scenario tail concentration:** construct scenarios using only the worst quantile of past outcomes within each bucket to test whether CVaR constraints still prevent oversized exposure.

These stress tests complement the standard validation window by directly probing the failure modes of structure-based signals under abrupt moves and degraded liquidity. The key stability criterion is that exposure remains bounded (by $W_{\max}$ and CVaR) and that drawdowns do not become dominated by clustered tail losses even when costs widen.

Table 9: Stress-test outcomes (hybrid system)

Stress condition	Return (%)	Sharpe	MDD (%)	$ES_{0.95}$
Extreme-bar subset (top 1% moves)	21.4	0.63	9.8	0.061
Cost widening ($c \times 2$)	38.2	0.99	7.4	0.048
Cost widening ($c \times 3$)	31.5	0.83	8.1	0.052
Tail-only scenarios	27.6	0.74	8.9	0.058

4.8 LP solve-time benchmark

To support the claim of real-time compatibility, we report solve times over the full backtest horizon for the chosen solver and typical S.

Table 10: LP solve-time benchmark (solver: HiGHS 1.x via `highspy` on Google Colab CPU)

Setting	Mean (ms)	Median (ms)	p95 (ms)	Max (ms)
Hybrid LP (chosen S)	1.9	1.6	3.8	6.4
LP-only (chosen S)	1.4	1.2	2.9	4.8

5 Discussion

The empirical results highlight that the performance advantage of the proposed hybrid framework does not arise from superior directional forecasting alone, but from the structured interaction between context-aware probabilities and explicit risk control. Calibrated machine-learning outputs provide a quantitative assessment of the likelihood that a given SMC/ICT setup will resolve favorably, while the linear-programming layer translates this assessment into position sizes that respect turnover, exposure, and tail-risk constraints.

Rules-only SMC/ICT strategies are effective at identifying regions of interest, particularly around liquidity sweeps and returns into imbalance zones, but their fixed or heuristic sizing schemes expose them to inconsistent risk. Failed mitigations and partial fills around order blocks generate clustered losses that dominate the left tail of the return distribution. This behavior is visible in both higher realized Expected Shortfall and larger maximum drawdowns relative to the hybrid system.

Machine-learning models used in isolation improve signal discrimination but lack structural grounding. Without explicit liquidity context or session conditioning, ML-only systems tend to trigger trades during statistically weaker environments, such as low-liquidity hours or compressed volatility regimes. The resulting increase in trade frequency raises transaction costs and dilutes expectancy, even when predictive accuracy is comparable. These findings reinforce the view that, in short-horizon trading, machine learning is more effective as a probabilistic filter than as a standalone decision engine.

The LP-only allocator exhibits strong robustness properties due to its reliance on CVaR constraints and turnover limits, but its use of naive expected edges prevents it from exploiting recurrent structural patterns. As a result, it under-allocates risk during high-quality opportunities and produces conservative but suboptimal performance. The hybrid framework resolves this limitation by injecting calibrated, structure-conditioned probabilities into the optimization, allowing the allocator to scale exposure selectively when the statistical and contextual evidence aligns.

Relative to structure-only heuristics, and as a comparison to related strands (Table 1), the proposed system adds (i) a calibrated probability layer for confidence-aware filtering and (ii) an explicit optimization layer for auditable risk control. Relative to ML-only prediction, it anchors features and targets in interpretable structural events and reduces probability miscalibration through post-hoc calibration. Relative to LP/CVaR allocation frameworks, it provides a data-driven edge proxy conditioned on session and regime rather than a naive constant edge, improving opportunity selection without sacrificing convex risk control.

While for generalization to other assets and timeframes, the proposed pipeline is modular: structural extractors can be applied to other liquid assets (major FX pairs, index futures, other commodities) and other timeframes provided that (i) the swing definition and displacement thresholds are re-scaled to the asset's volatility and tick size, (ii) session definitions reflect the asset's liquidity cycle, and (iii) cost modeling is adapted to the venue. Higher-frequency data can improve timing and mitigate OHLC path ambiguity, but also increases microstructure noise and sensitivity to spread/latency, in that setting, tighter turnover constraints and more granular scenario buckets are typically required.

5.1 External benchmark positioning (SOTA context)

Direct comparison across algorithmic trading studies is inherently limited due to differences in asset universes (single-asset vs. multi-asset), sampling frequency, leverage constraints, transaction-cost models, and evaluation windows. Nevertheless, to position the magnitude of our risk-adjusted performance in the context of recent ML+optimization or risk-constrained allocation literature, Table 11 reports selected, numerically stated benchmarks (Sharpe and/or maximum drawdown when available) and contrasts them with our *development-period* trading metrics (Table 5). Out-of-sample evaluation in this paper is reported via calibration and downside-risk diagnostics on the held-out window (Table 6).

Our single-asset XAUUSD setting differs from multi-asset allocation frameworks, therefore, the intent is not to claim strict dominance, but to show that (i) our achieved risk-adjusted performance is within the range reported by modern risk-aware ML allocation frameworks, and (ii) our

drawdown control is consistent with the design goal of tail-risk constrained sizing under cost-aware execution assumptions.

Table 11: Numerical positioning versus representative ML + risk/optimization benchmarks. Values are taken as reported in the cited works; settings are not directly comparable. For this work, Sharpe and Max DD correspond to the *development period* (Table 5); out-of-sample evaluation is reported via diagnostics in Table 6

Reference	Market / assets	Core method	Sharpe	Max DD	Notes (key differences)
This work	XAUUSD (single asset)	Calibrated ML + LP (CVaR) sizing	**1.18**	**6.8%**	**Dev** (Jan 2023–May 2025); session/regime scenarios; explicit costs; sizing constrained by CVaR
Agal et al. (2025) [32]	Multi-asset allocation	ML-driven risk-based allocation	1.38	16.2%	Portfolio setting; different universe/costs; focuses on allocation under regime adaptation
Moung et al. (2024) [33]	FX (multiple pairs)	RL trading framework	0.16–0.70	–	Sharpe reported; drawdown not consistently stated in the venue summary

Differences in leverage, sampling frequency, universe breadth, and transaction-cost definitions prevent apples-to-apples ranking, the table is provided only to contextualize numerical magnitude.

While exact comparisons across studies are limited by differing assets, horizons, and cost assumptions, the *development-period* risk-adjusted magnitude we obtain is within reported ranges for short-horizon ML trading on liquid FX/commodities under realistic frictions. Out-of-sample, we focus on probability reliability and downside-risk diagnostics (Table 6), which remain stable and support the robustness of the calibration and risk-control layers.

The main contribution is mechanistic: calibration mitigates overconfident sizing, and LP with CVaR constraints translates forecast confidence into bounded exposure that improves tail stability versus ML-only sizing.

Limitations

This work relies on hourly OHLC data and therefore cannot fully observe intrabar path risk (e.g., stop-outs within the bar) or order-book dynamics. The structural definitions (FVG/OB/breaks) are sensitive to news-driven discontinuities where gap risk dominates and liquidity thins, the stress tests in Section 4 aim to quantify this vulnerability but cannot eliminate it. The approach is also domain-specific: features are crafted around SMC/ICT primitives and may not transfer to assets without comparable intraday liquidity cycles. Finally, while walk-forward evaluation reduces overfitting, any parameterization (swing detection, thresholds, bucket boundaries) still induces a search space that should be kept small and audited.

6 Conclusion

This study proposes a reproducible framework that bridges Smart Money Concepts and ICT-style price action with modern machine learning and operations research. By formalizing SMC/ICT primitives as deterministic, leakage-safe extractors, learning calibrated probabilities for monetizable structural outcomes, and embedding these probabilities into a linear-programming allocator with CVaR constraints, the framework transforms qualitative trading ideas into a risk-aware decision system.

Empirical evaluation on two years of hourly XAUUSD data, complemented by a fully held-out validation window, demonstrates that the hybrid system consistently outperforms rules-only, ML-only, and LP-only baselines on a risk-adjusted basis under realistic transaction costs and turnover limits. Performance gains are primarily attributable to improved tail-risk control and selective scaling of exposure during high-quality setups, rather than to increased trade frequency or optimistic assumptions.

Beyond the specific asset studied, the proposed formulation is intentionally transparent and modular. Each component can be independently audited, replaced, or extended, making the approach suitable for multi-asset portfolios, alternative timeframes, or richer liquidity features when available. Future work may explore portfolio-level optimization, regime-dependent constraints, and the integration of higher-frequency or event-based data to better capture intrabar path risk.

Perspectives Promising extensions include (i) multi-asset and multi-timeframe joint allocation, (ii) more realistic execution modeling (spread variation, slippage, and latency) embedded as linear constraints, (iii) robust or distributionally robust variants of the CVaR constraint to reduce regime sensitivity, and (iv) online recalibration and drift monitoring for the probabilistic layer.

In summary, the results support the view that combining domain-specific structure, calibrated probabilistic inference, and linear risk-aware allocation offers a viable and robust pathway for systematizing discretionary price-action methodologies within a rigorous quantitative framework.

References

[1] J. Hasbrouck. *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press, 2007.

[2] J.-P. Bouchaud, J. D. Farmer, and F. Lillo. How Markets Slowly Digest Changes in Supply and Demand. In T. Hens and K. R. Schenk-Hoppé (eds.), *Handbook of Financial Markets: Dynamics and Evolution*, pp. 57–160. Elsevier, 2009.

[3] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. doi:10.1023/A:1010933404324.

[4] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 785–794, 2016. doi:10.1145/2939672.2939785.

[5] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. doi:10.1162/neco.1997.9.8.1735.

[6] M. H. Oukhouya, N. Angour, N. Aboutabit, and I. Hafidi. Comparative Analysis of ARDL, LSTM, and XGBoost Models for Forecasting the Moroccan Stock Market During the COVID-19 Pandemic. *Informatica (Slovenia)*, 49(14):203–214, 2025. doi:10.31449/inf.v49i14.5751.

[7] R. T. Rockafellar and S. Uryasev. Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2(3):21–41, 2000.

[8] R. T. Rockafellar and S. Uryasev. Conditional Value-at-Risk for General Loss Distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002. doi:10.1016/S0378-4266(02)00271-6.

[9] R. Almgren and N. Chriss. Optimal Execution of Portfolio Transactions. *Journal of Risk*, 3(2):5–39, 2001.

[10] C. L. Osler. Currency Orders and Exchange-Rate Dynamics: An Explanation for the Predictive Success of Technical Analysis. *The Journal of Finance*, 58(5):1791–1819, 2003. doi:10.1111/1540-6261.00588.

[11] C. L. Osler. Stop-Loss Orders and Price Cascades in Currency Markets. *Journal of International Money and Finance*, 24(2):219–241, 2005. doi:10.1016/j.jimonfin.2004.12.002.

[12] A. S. Kyle. Continuous Auctions and Insider Trading. *Econometrica*, 53(6):1315–1335, 1985. doi:10.2307/1913210.

[13] M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison. Limit Order Books. *Quantitative Finance*, 13(11):1709–1742, 2013. doi:10.1080/14697688.2013.803148.

[14] R. Cont, A. Kukanov, and S. Stoikov. The Price Impact of Order Book Events. *Journal of Financial Econometrics*, 12(1):47–88, 2014. doi:10.1093/jjfinec/nbt003.

[15] A. W. Lo, H. Mamaysky, and J. Wang. Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation. *The Journal of Finance*, 55(4):1705–1765, 2000. doi:10.1111/0022-1082.00265.

[16] W. Brock, J. Lakonishok, and B. LeBaron. Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *The Journal of Finance*, 47(5):1731–1764, 1992. doi:10.1111/j.1540-6261.1992.tb04681.x.

[17] R. Sullivan, A. Timmermann, and H. White. Data-Snooping, Technical Trading Rule Performance, and the Bootstrap. *The Journal of Finance*, 54(5):1647–1691, 1999.

[18] H. White. A Reality Check for Data Snooping. *Econometrica*, 68(5):1097–1126, 2000.

[19] T. Fischer and C. Krauss. Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions. *European Journal of Operational Research*, 270(2):654–669, 2018. doi:10.1016/j.ejor.2017.11.054.

[20] S. Gu, B. Kelly, and D. Xiu. Empirical Asset Pricing via Machine Learning. *Review of Financial Studies*, 33(5):2223–2273, 2020. doi:10.1093/rfs/hhaa009.

[21] A. Niculescu-Mizil and R. Caruana. Predicting Good Probabilities with Supervised Learning. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005. doi:10.1145/1102351.1102430.

[22] B. Zadrozny and C. Elkan. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2002. doi:10.1145/775047.775151.

[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of ICML*, 2017. arXiv:1706.04599.

[24] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent Measures of Risk. *Mathematical Finance*, 9(3):203–228, 1999. doi:10.1111/1467-9965.00068.

[25] H. Konno and H. Yamazaki. Mean-Absolute Deviation Portfolio Optimization Model and Its Applications to Tokyo Stock Market. *Management Science*, 37(5):519–531, 1991. doi:10.1287/mnsc.37.5.519.

[26] P. Krokhmal, J. Palmquist, and S. Uryasev. Portfolio Optimization with Conditional Value-at-Risk Objective and Constraints. *Journal of Risk*, 4(2):11–27, 2002.

[27] D. H. Bailey and M. López de Prado. The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality. *The Journal of Portfolio Management*, 40(5):94–107, 2014. doi:10.3905/jpm.2014.40.5.094.

[28] D. H. Bailey, J. M. Borwein, M. López de Prado, and Q. J. Zhu. The Probability of Backtest Overfitting. *Journal of Computational Finance*, 20(4):39–69, 2017. Preprint available at SSRN:2326253.

[29] G. W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[30] J. C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

[31] M. Kull, T. M. Silva Filho, and P. Flach. Beta Calibration: a Well-Founded and Easily Implemented Improvement on Logistic Calibration for Binary Classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[32] S. Agal, K. Raulji, and N. D. Odedra, "A machine learning approach to risk based asset allocation in portfolio optimization," *Scientific Reports*, vol. 15, 42263, 2025. doi:10.1038/s41598-025-26337-x.

[33] E. G. Moung *et al*, "Optimizing foreign exchange trading performance through reinforcement machine learning framework," in *14th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2024. doi:10.1109/ICCKE65377.2024.10874712.

[34] F. D. Foster and S. Viswanathan. Variations in Trading Volume, Return Volatility, and Trading Costs: Evidence on Recent Price Formation Models. *The Journal of Finance*, 48(1):187–211, 1993. DOI: 10.1111/j.1540-6261.1993.tb04706.x.