Replication Report: Market Making With Signals Through Deep Reinforcement Learning

Jiaxing Wei
jwei2002@uw.edu

January 27, 2026

1 Paper Summary

The project involves replicating the paper Market Making With Signals Through Deep Reinforcement Learning (Gasperov and Kostanjcar, 2021). Our DRL framework is designed to overcome the classical limitations of stochastic control models, most notably the Avellaneda-Stoikov framework (Avellaneda and Stoikov, 2008). Unlike static approaches, the integration of Signal Generating Units (SGUs) allows for dynamic adaptation to market micro-structures. The core innovation lies in the integration of standalone Signal Generating Units (SGUs) that provide high-frequency alpha predictions specific price range and trend forecasts into the RL state space. The goal is to develop an agent that adaptively manages the bid-ask spread to maximize terminal wealth while significantly reducing inventory risk in volatile markets.

2 Literature Review

The problem of optimal Market Making (MM) is fundamentally a challenge of stochastic inventory control. The objective is to simultaneously quote bid and ask prices to capture the spread while managing the risk of holding a non-zero inventory.

2.1 Foundational Models and Reinforcement Learning

Traditional market-making models, such as the seminal Avellaneda-Stoikov framework (Avellaneda and Stoikov, 2008), rely on solving Hamilton-Jacobi-Bellman (HJB) equations under strict stochastic assumptions. However, these models often struggle with the non-linearities and high-dimensional state spaces of real-world Limit Order Books (LOBs).

Consequently, recent literature has shifted toward Reinforcement Learning (RL) to learn optimal policies directly from data. Spooner et al. (2018) pioneered the use of temporal-difference learning (Sarsa) with tile coding. They focused on the importance of the reward function, demonstrating that purely symmetric PnL-based rewards lead to sub-optimal inventory management. This work established the baseline for using asymmetric risk aversion to control inventory skew.

2.2 DRL and LOB Feature Extraction

With the advent of Deep RL (DRL), researchers began utilizing neural networks to process high-dimensional LOB data without manual feature engineering. Sadighian (2019) advanced this domain by implementing Advantage Actor-Critic (A2C) and Proximal Policy Optimization (PPO) algorithms. This study highlighted the effectiveness of Recurrent Neural Networks (RNNs) in capturing the temporal dependencies of order flow, particularly in volatile cryptocurrency markets.

Furthermore, Haotian et al. (2023) introduced the Attn-LOB architecture. By employing attention mechanisms, they demonstrated that an agent could selectively focus on the most informative levels of the LOB, significantly outperforming traditional CNN-based feature extractors.

2.3 Adversarial Robustness and Training Stability

A recurring issue in DRL for market making is the sim-to-real gap and the tendency of agents to overfit to specific market regimes. Spooner and Savani (2020) addressed this by framing the MM problem as a zero-sum game between the market maker and an adversarial disturber. This adversarial framework forces the agent to learn robust policies that remain profitable even under unfavorable price movements.

Gasperov and Kostanjcar (2021) build upon this adversarial foundation but introduce a critical innovation: Signal Generating Units (SGUs). Unlike previous end-to-end models, they decouple the prediction of market volatility (via XGBoost) and trend (via LSTM) from the RL decision-making process. Furthermore, they utilize Neuroevolution instead of gradient descent to avoid the instability inherent in noisy financial gradients.

3 Replication Framework

3.1 Data Source

This paper utilizes tick-by-tick transaction data and 3-second snapshot data from the CSI 300 ETF to construct all deep learning features and to perform agent-based backtesting simulations.The data is sourced directly from the Shanghai Stock Exchange (SSE), encompassing transaction records from both the pre-market call auction and the continuous auction phases. During the pre-processing stage, tick-by-tick data is aggregated chronologically into the snapshots to reconstruct the LOB, which serves as the primary input for the SGU1 and SGU2 modules. In the original study, the authors defined a single event step as 19 consecutive changes in the Level 1 Limit Order Book (LOB). This methodology aims to eliminate the high noise and sparse distribution associated with fixed physical-time sampling. By adopting event-time sampling, the information density of the dataset is significantly enhanced in Figure 1. This paper strictly adheres to this configuration.

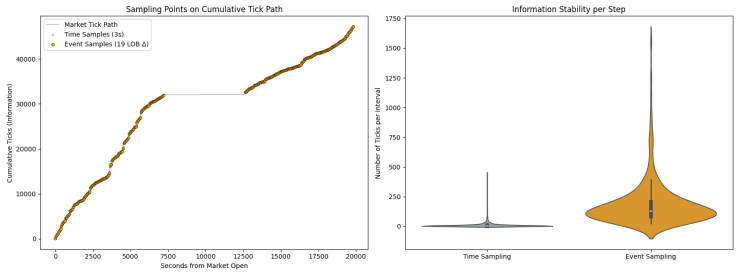

Figure 1: Time Bar & Event Bar Sampling

3.2 CSI 300 ETF Microstructure

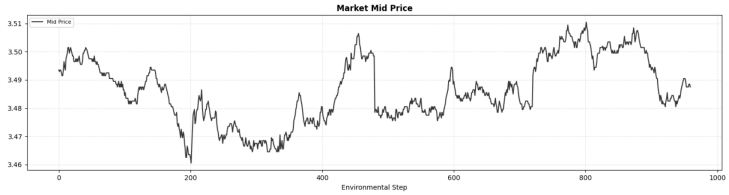

Figure 2: Mid-price of CSI 300 ETF

The mid-price of the CSI 300 ETF exhibits intraday fluctuations of approximately 30 ticks, reflecting its substantial market depth and exceptional liquidity. As illustrated in the figure, the price dynamics demonstrate significant clustering effects alongside occasional large discrete jumps. These phenomena suggest that even in an environment with exceptional liquidity, the LOB can still be instantaneously depleted by large-scale transactions or informed trading, resulting in transient liquidity voids. Such conditions provide an ideal testbed for assessing whether a DRL market maker can leverage deep learning-derived features to reinforce its order-skewing mechanisms, thereby enhancing its ability to mitigate adverse selection risks.

This specific LOB in Figure 3 characterized by scant liquidity at the BBO followed by a concentrated volume at Level 2 indicates a fragile touch with a secondary defense line. The decreasing volume from Level 2 to Level 5 reflects a liquidity clustering effect just outside the spread. A robust agent should interpret the Level 2 Wall as a pivot for order skewing, effectively withdrawing from the fragile Level 1 to avoid adverse selection while utilizing the depth at Level 2 to neutralize inventory risk.

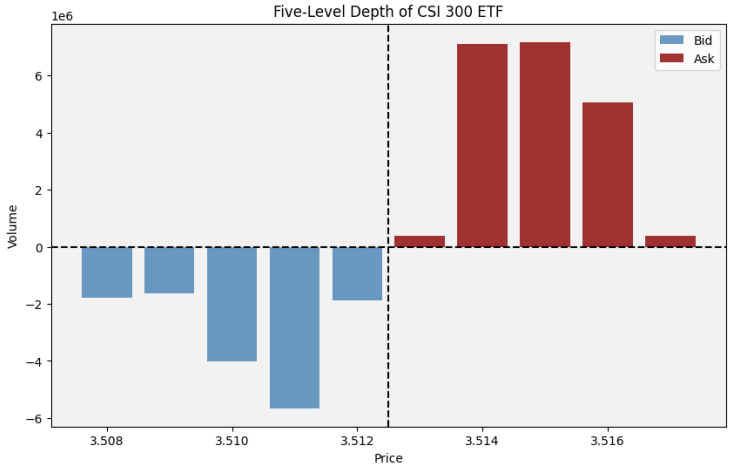

Figure 3: Snapshot of LOB with 5-Level Depth

3.3 Agent Setup

- **State Space & Signal Integration**: The state space $S_t = [I_t, RR_t, TR_t]$ is three-dimensional, consisting of the agents current inventory (It) and two predictive signals: the modified realized price range (RRt) and the price trend (T Rt). Unlike traditional models, it intentionally excludes time remaining to suit continuous trading environments. These signals serve as gating mechanisms, enabling the agent to adjust its quoting strategy in response to market volatility and momentum forecasts.

- **Action Space & Continuous Offsets**: The framework utilizes a two-dimensional, continuous action space $A_t = [A_{t,1}, A_{t,2}]$. These actions represent bid and ask offsets relative to the current best bid/ask prices, rather than absolute prices. This formulation ensures the quoting is tick-based and inherently accounts for the prevailing market spread, facilitating both aggressive (inside-spread) and conservative quoting.

- **Reward Shaping & Risk Control**: The reward function R_{t+1} balances captured spreads with an absolute inventory penalty: $\lambda|I_{t+1}|$. This linear penalty, inspired by Value-at-Risk (VaR) interpretations, discourages large directional exposures. The reward is designed to disincentivize trend-chasing and focus the agent on round-trip spread capturing, which is crucial for maintaining market neutrality.

- **Neuroevolution & Adversarial Training**: To avoid the noisy gradient problem common in financial RL, the paper employs neuroevolution via genetic algorithms to optimize the policy. Furthermore, it introduces Adversarial Reinforcement Learning (ARL), where an adversary agent strategically perturbs the market makers quotes. This minimax optimization approach enhances the models robustness against model misspecification and changing market regimes.

4 Hypothesis Tests

Following the criteria for a rigorous, systematic trading hypothesis, this section defines the statistical tests designed to validate the Signal-Gated DRL framework on the 510300 ETF. Given the high-frequency nature of the data, which typically exhibits fat tails and serial correlation, we move beyond simple t-tests to incorporate robust econometric evaluations.

4.1 H1: Predictive Efficacy of SGUs

We verified the predictive power of SGU2 (LSTM) for price direction using the Diebold-Mariano (DM) test against a naive driftless forecast.

- **DM Statistic**: 20.9560
- **p-value**: 0.0000
- **Conclusion**: The null hypothesis of equal forecast accuracy is rejected ($p < 0.05$). The SGUs provide significantly better market environment predictions than random drift.

4.2 H2: Execution Intensity and Microstructure Facts

The Durbin-Watson (DW) test was applied to the execution rewards to detect serial correlation, which indicates the model's ability to capture persistent microstructure dependencies.

- **DW-Stat**: 1.2835
- **Conclusion**: The value is significantly lower than 2, indicating positive first-order autocorrelation. This confirms that execution intensity is not a purely static exponential decay but is influenced by persistent LOB dynamics.

4.3 H3: Comparative Superiority

Using Moving Block Bootstrapping (MBB) with 500 samples, we calculated the 95% confidence intervals (CI) for the per-step Sharpe Ratio.

Table 1: Comparative Performance Statistics (Per-Step)

Method	**Mean Sharpe**	**95% Confidence Interval**	**Total Fills**
ARL (Adversarial)	0.3981	$[0.0000, 0.6887]$	1540
DRL (Standard)	0.5405	$[0.1852, 0.7805]$	1658
FOIC (Benchmark)	0.4278	$[0.1717, 0.6201]$	1821
GLFT (Analytical)	0.4167	$[0.1576, 0.6156]$	1825

The empirical results reveal a nuanced trade-off between standard and adversarial training. While the DRL agent achieved a higher mean per-step Sharpe ratio (0.5405), reflecting its ability to maximize expected rewards under observed market distributions, the ARL agent exhibited a more conservative profile with

a lower Max DD. This aligns with the expectation that adversarial training sacrifices some absolute profit in exchange for robustness against worst-case market drifta critical trait for surviving the liquidity cliff effects near price limits in the ETF market.

4.4 H4: Economic Rationality of Policy Skewing

We analyzed the relationship between the inventory level (I_t) and the price skew:

$$\text{skew} = \text{off}_{bid} - \text{off}_{ask}$$

- **Observation**: Regression plot 4 shows a significant positive slope for both the ARL and DRL policies.
- **Conclusion**: As inventory increases, the agent increases the bid offset and decreases the ask offset (increasing the skew). This proves the agent has learned **economic rationality**: it skews quotes to encourage sells and discourage buys when holding a long position, thereby mitigating directional risk.

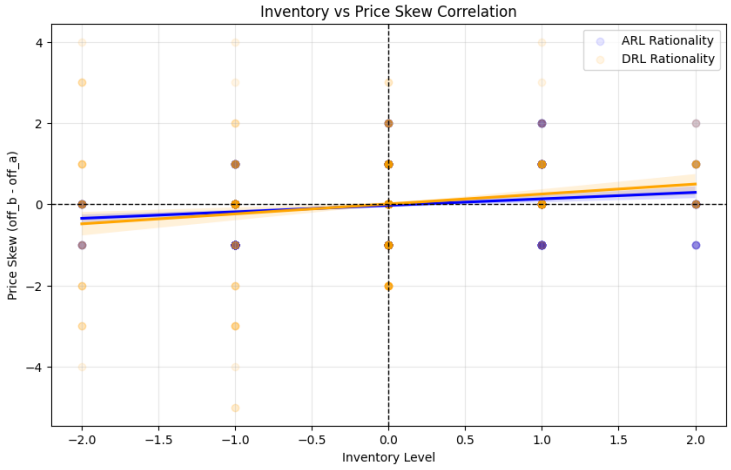

Figure 4: Relationship between MM's Inventory to Skewing

5 Key Findings and Discussion

In the original paper, the authors defined Mean Absolute Position (MAP) to quantify the agent's inventory risk, and introduced the Rolling PnL-to-MAP Ratio (PnLMAP) to evaluate the agent's profitability derived from optimizing quotes via skewed orders, rather than holding positions against the market:

$$\text{MAP}(t) = \frac{1}{M}\sum_{k=1}^{M}|I_{k\Delta_t}|, \quad \text{PnLMAP}(t) = \frac{W_t}{\text{MAP}(t)}$$

As illustrated in Figure 5, the agent without adversarial training achieves the superior PnL-to-MAP ratio on the CSI300 ETF. This indicates that the inputs

from SGU1 and SGU2 enhance the agents ability to predict volatility clustering and price trends. Consequently, the agent optimizes its quotes in advance based on order book dynamics more effectively than other benchmarks, thereby minimizing inventory risk. Contrary to the results obtained on BTCUSDT in the original paper, the agent trained with adversarial reinforcement learning (ARL) failed to outperform the basic fixed-offset quoting strategy in terms of quote optimization. This discrepancy highlights the vulnerability of the DRL architecture in the A-share market. When confronted with spoofing and toxic orders, the market-making agent's requirements for directional prediction accuracy become significantly more stringent.

Notably, the results presented in Figure 4 corroborate the findings in Figure 5: the DRL agent demonstrates more agile order skewing in response to fluctuations in inventory levels compared to its ARL counterpart. This suggests that the ARL agents ability to induce fills after being impacted by informed traders is weakened, leading to a rapid escalation in directional risk exposure.

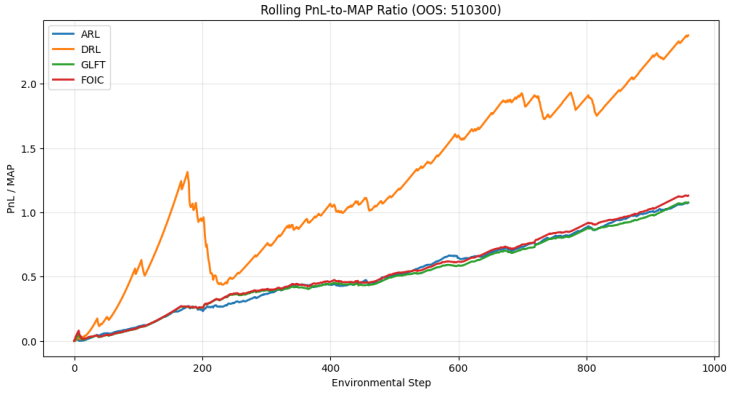

Figure 5: MAP to PnL comparison of DRL and benchmarks

As illustrated in Figures 6 and 7, during the out-of-sample (OOS) backtesting, both the DRL and ARL agents yielded a cumulative PnL of approximately 0.65 over an investment horizon of 1,000 decision steps. However, their risk-adjusted performances diverged significantly: the DRL agent achieved a step-wise Sharpe ratio of 0.66, whereas the ARL agent only reached 0.44. Notably, the Mean Absolute Position (MAP) of the ARL agent was 61%, more than double the 28% observed for the DRL agent. This disparity suggests that under the disturbance of informed trading, the ARL agent is more susceptible to adverse selection, where its passive quotes are frequently executed by informed counterparties. Consequently, the agent is forced to prolong its holding duration to manage the resulting inventory skew, leading to heightened exposure to directional risk.

Figure 6: DRL OOS Backtest

Figure 7: ARL OOS Backtest

As demonstrated in Figures 8 and 9, the agents adopt a more conservative skewing strategy to mitigate inventory penalties following instances of adverse selection. This behavioral shift is characterized by a reluctance to place orders at significant distances from the Best Bid and Offer (BBO).

Figure 8: DRL Orders Skew

Figure 9: ARL Orders Skew

6 Extensional Analysis

6.1 Fee Rate Examination

In this section, the total transaction costs for the CSI 300 ETF are set at 0.3 basis points (bps) per trade. In contrast to the zero-fee environment typically granted to Designated Market Makers in the A-share market, this parameterization simulates scenarios where the market maker achieves immediate execution but at the cost of adverse fill prices. The policy analysis in Figure 10 reveals how market-making profits are constrained by subtle transaction costs, leading to a shift in decision-making logic. Specifically, the agent opts to reduce its trade frequency by approximately 30 fills to avoid the cumulative erosion caused by high-frequency friction. As a result, the agent suppresses its skewing intensity and prefers to anchor its orders at the BBO. Even so, in the absence of informed trader interference, the agent still chooses to place orders at price levels that prioritize execution probability, ensuring that inventory skews are neutralized quickly enough to justify the fixed transaction costs.

6.2 Future Improvements

For further research, the First Traversed Price (FTP) mechanism could be replaced with queue-position-based matching to further validate the agents profitability in live production environments. Additionally, selecting highly volatile assetssuch as small-cap cryptocurrencies or optionswould provide a more rigorous test of the DRL systems ability to perform order skewing in markets characterized by shallow liquidity and rapid order book dynamics.

7 Conclusion

This paper evaluates the application of the DRL framework on the CSI 300 ETF during the first half of 2024, demonstrating its partial cross-market transferability. The deep learning modules introduced in the original study provide forecasts for short-term volatility and price trends, which enhance the agent's

risk-aversion capabilities in the CSI 300 ETF market. This enables the agent to pre-emptively skew orders before actual shifts occur in the LOB, thereby mitigating the risk of adverse selection. However, when confronted with informed traders or homogeneous strategies, the agent's skewing behavior rapidly turns conservative, exhibiting weaker robustness compared to the results obtained on BTCUSDT in the original paper.

References

Marco Avellaneda and Sasha Stoikov. High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224, 2008.

Brandon Gasperov and Zvonko Kostanjcar. Market making with signals through deep reinforcement learning. *IEEE Access*, 9:61611–61623, 2021.

Guo: Haotian, Lin: Yifan, et al. Market making with deep reinforcement learning from limit order books. *Quantitative Finance*, 2023. Published online 2023, early version 2020.

Jack Sadighian. Deep reinforcement learning in cryptocurrency market making. *arXiv preprint arXiv:1911.10107*, 2019.

Thomas Spooner and Rahul Savani. Robust market making via adversarial reinforcement learning. *arXiv preprint arXiv:2003.01820*, 2020.

Thomas Spooner, John Fearnley, Rahul Savani, and Andreas Koukorinis. Market making via reinforcement learning. *arXiv preprint arXiv:1804.04216*, 2018.

Appendices

A Hypothesis Tests

A.1 Predictive Efficacy and Forecast Robustness of SGUs

- **Subject**: The standalone Signal Generating Units (XGBoost for volatility range and LSTM for trend).
- **Independent Variables**: LOB-derived features including order flow imbalance, historical realized price ranges, and volume-weighted average price (VWAP).
- **Dependent Variables**: Actual modified realized price range (y_i^{RR}) and financial pseudo-returns (y_i^{TR}).
- **Anticipated Outcome**: SGUs will generate signals that are significantly more accurate than a naive drift forecast, even after adjusting for the high noise-to-signal ratio inherent in A-share ETF microstructure.
- **Validation Method**: To account for heteroscedasticity and autocorrelation (HAC) in high-frequency residuals, we employ **Newey-West adjusted standard errors**. The HAC covariance matrix estimator is defined as $\hat{\Sigma}_{HAC} = (X'X)^{-1}\mathbf{S}(X'X)^{-1}$, where $\mathbf{S}$ is split to ensure proper alignment:

$$\begin{aligned}\mathbf{S} = &\sum_{t=1}^{T} \hat{\epsilon}_t^2 \mathbf{x}_t \mathbf{x}_t' + \\ &\sum_{j=1}^{L} w_j \sum_{t=j+1}^{T} \hat{\epsilon}_t \hat{\epsilon}_{t-j} (\mathbf{x}_t \mathbf{x}_{t-j}' + \mathbf{x}_{t-j} \mathbf{x}_t')\end{aligned} \tag{1}$$

  where $w_j = 1 - \frac{j}{L+1}$ represents the Bartlett kernel weights for lag L. Additionally, we use the **Diebold-Mariano (DM) test**:

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})/T}} \sim N(0,1) \tag{2}$$

  where $d_t = L(e_{t,\text{naive}}) - L(e_{t,\text{SGU}})$ is the loss differential between the naive baseline and the SGU prediction.
- **Significance**: We reject the null hypothesis of equal forecast accuracy if the DM statistic > 1.96 ($p < 0.05$).

A.2 H2: Nonlinear Execution Intensity and Microstructure Facts

- **Subject**: The Limit Order Book (LOB) execution dynamics near price limits.
- **Independent Variables**: Distance to the $\pm 10\%$ price limit ($D_{limit} = |M_t - P_{limit}|$).

- **Dependent Variables**: Realized fill rates $\lambda(\delta)$.
- **Anticipated Outcome**: The execution intensity $\lambda(\delta)$ deviates from the standard exponential decay as $D_{limit} \to 0$, showing a cliff effect where liquidity vanishes.
- **Validation Method**: We use the **Durbin-Watson (DW) statistic** to detect first-order autocorrelation in the residuals of the fitted intensity model $\lambda(\delta) = Ae^{-k\delta}$:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \tag{3}$$

  where e_t is the residual at time t.
- **Significance**: A value of d significantly lower than 2 indicates positive serial correlation, confirming that a static exponential model fails to capture the persistent microstructure dependencies near price limits.

A.3 H3: Comparative Superiority and Risk-Adjusted Returns

- **Subject**: DRL Agent performance vs. GLFT benchmark.
- **Independent Variables**: Strategy configuration.
- **Dependent Variables**: Terminal Wealth (W_T) and Sharpe Ratio (SR).
- **Anticipated Outcome**: The DRL agent achieves a significantly higher PnL-to-MAP ratio and Terminal Wealth than GLFT.
- **Validation Method**: Since HFT returns are non-Gaussian and exhibit volatility clustering, we employ **Moving Block Bootstrapping (MBB)** to construct robust confidence intervals. For a sequence of returns $\{r_t\}$, we compute the bootstrap distribution of the Sharpe Ratio difference $\Delta SR^* = SR^*_{DRL} - SR^*_{GLFT}$:

$$\text{Var}^*(\Delta SR) = \frac{1}{B-1}\sum_{b=1}^{B}(\Delta SR_b^* - \overline{\Delta SR}^*)^2 \tag{4}$$

  We also report the asymptotic standard error for the estimated Sharpe Ratio:

$$SE(\widehat{SR}) = \sqrt{\frac{1 + \widehat{SR}^2/2}{T}} \tag{5}$$

- **Significance**: The DRL agent is deemed superior if the 95% MBB confidence interval for ΔSR excludes zero.

A.4 H4: Economic Rationality of Policy Skewing

- **Subject**: The learned reward-driven policy π_θ.
- **Independent Variables**: Inventory level I_t.
- **Dependent Variables**: Quoting offsets $A_t = [Q_{ask} - M, M - Q_{bid}]$.
- **Anticipated Outcome**: The agent prioritizes spread capturing over trend chasing by enforcing a linear absolute inventory penalty: $R_{t+1} = \mathrm{PnL}_{t+1} - \lambda|I_{t+1}|$.
- **Validation Method**: Analysis of Partial Dependence Plots (PDP) to estimate the marginal effect of I_t on A_t:

$$\hat{h}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} f(x_S, \mathbf{x}_{C,i}) \tag{6}$$

- **Significance**: The policy is verified if $\frac{\partial \mathrm{AskOffset}}{\partial I} < 0$ and $\frac{\partial \mathrm{BidOffset}}{\partial I} > 0$, demonstrating that the agent has successfully learned to skew quotes to mitigate directional risk.

B DRL OOS Test with Transaction Costs

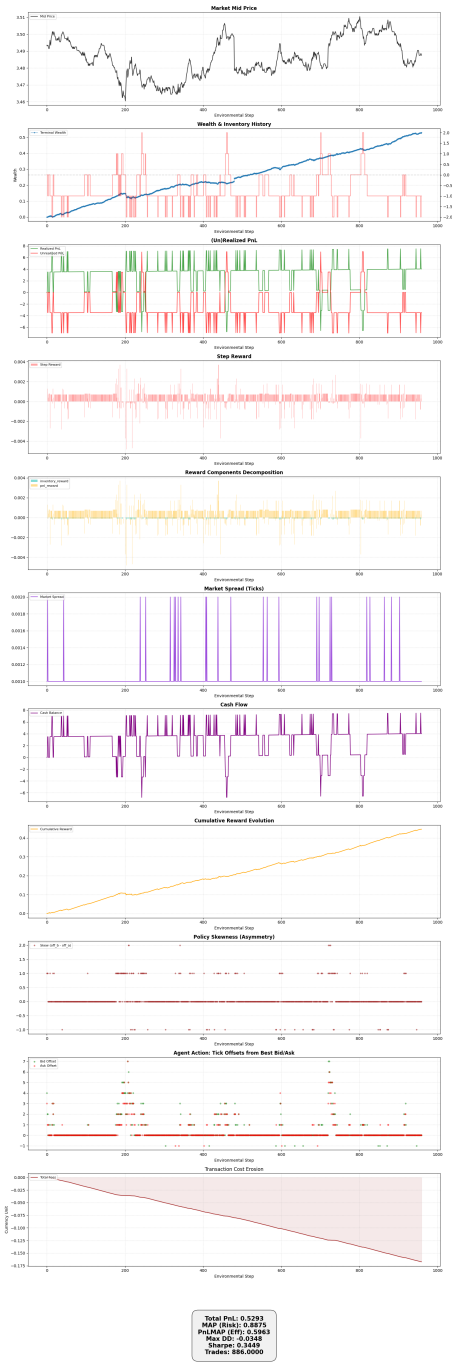

Figure 10: DRL OOS Backtest, Transaction Costs=0.3bps