

Day 4: First Learning Policy

What behavior emerged

The agent developed a non-random, conservative trading policy. Over training, actions shifted away from uniform randomness and toward structured behavior, with the agent selectively buying and selling rather than acting arbitrarily or freezing into a hold-only policy. Inventory remained bounded, indicating that the agent learned to manage exposure instead of accumulating unchecked positions.

What the reward is incentivizing

The reward function primarily incentivizes stability and risk control over profit maximization. Incremental PnL provides a learning signal, but the inventory and risk penalties strongly discourage large positions and volatile portfolio changes. As a result, the agent learns to prefer low-risk actions that avoid drawdowns, even if that limits upside gains.

One concrete flaw to fix next

The reward signal is too symmetric around zero PnL, which encourages the agent to converge toward near-zero-return behavior. A concrete next fix is to increase the relative weight of positive PnL (or reduce the inventory penalty slightly) so that profitable trades are more clearly distinguished from neutral actions, without removing risk awareness.