Day 5: Learnability Verdict

Does the agent learn?

Yes, the agent demonstrates clear evidence of learning under sustained optimization. Mean episode reward improves from strongly negative values during early training and stabilizes near zero over time, indicating that the agent learns to avoid systematically loss-making behavior. Policy entropy changes gradually rather than remaining flat or collapsing, showing that the agent forms preferences instead of acting randomly. This confirms that the environment, reward function and PPO algorithm together form a coherent learning system.

What behavior dominates?

The dominant behavior is risk-averse, low-exposure trading. The agent actively uses both buy and sell actions, with a mild bias toward selling, and does not fall into degenerate strategies such as always holding or always trading. The stabilized reward near zero suggests that the agents has learned that aggressive trading is penalized under the current reward structure and therefore prefers behavior that minimizes inventory risk and drawdowns rather than maximizing raw PnL.

One concrete fix if learning is weak

If learning needs to be strengthened, a concrete next step would be to reduce the risk penalty relative to incremental PnL, for example by lowering ‘k_risk’ or using a linear (instead of a quadratic) risk term. This would increase the signal to noise ratio of profitable actions, allowing the agent to more clearly distinguish between neutral and beneficial trades without removing risk awareness entirely.