Day 9: Hyperparameter Study v1

Which parameters matter most?

Entropy coefficient (ent_coeff)

Entropy coefficient emerged as the most influential hyperparameter.
High importance indicates that exploration exploitation balance strongly affects agent performance.
Very low entropy values led to premature convergence and brittle policies.
Moderate entropy values enabled continued exploration, allowing the agent to adapt to the stochastic momentum-driven price process.
In a non-stationary microstructure environment, how quickly the agent commits to a strategy matters more than the strategy itself.

Learning rate (learning_rate)

Learning rate was the second most important parameter, closely following entropy. If too high, there's unstable updates and degraded performance but if too low, it adapts slow and underutilizes the learning budget. The best performing trials consistently used low but non-minimal learning rates.
The environment produces noisy, high-variance gradients. Stable learning requires conservative step sizes.

Which hyperparameters barely matter?

Discount factor (gamma)

The discount factor showed relatively lower importance compared to entropy and learning rate. Performance was robust across a wide range of values for gamma and there was no sharp sensitivity near gamma tending to 1.0 or lower bounds.
The task is dominated by short to medium horizon dynamics. Long term credit assignment plays a limited role because inventory penalties and price momentum act locally in time

What this implies about the environment and award

Environment characteristics:

- The market simulation is well conditioned and symmetric.
- There is no guaranteed structural edge.
- Profits arise from timing and stability, not exploitation of bias.

Reward implications:

- Hyper parameter tuning improves behavioral quality (stability, adaptability).
- It does not fundamentally change the attainable profit frontier.
- Large performance swings across trials reflect learning dynamics, not reward hacking.

This confirms that the reward function is aligned, the environment is not degenerate and that the observed improvements are casual and reproducible.