1. Which algorithm converges fastest?

Policy Iteration converges the fastest because policy evaluation rapidly stabilizes in a deterministic environment, often in only a few iterations. Value Iteration typically requires more sweeps of updates before convergence. Q-learning is the slowest since it requires iterative sampling rather than full-sweep updates.

2. How does the discount factor affect value magnitude?

A higher discount factor ($\gamma = 0.9$) produces much larger value estimates because long-term rewards are weighted strongly. A moderate discount factor ($\gamma = 0.5$) reduces values since distant rewards matter less. A very low discount factor ($\gamma = 0.1$) yields small values because the agent focuses almost exclusively on immediate rewards.

3. How does γ influence preference for long-term vs short-term rewards?

With $\gamma = 0.9$, the agent heavily values distant future rewards and selects policies that optimize long-term outcomes. With $\gamma = 0.5$, long-term and short-term rewards are balanced. With $\gamma = 0.1$, only immediate outcomes matter and the agent becomes short-sighted.

4. How does γ affect policy behavior?

At $\gamma = 0.9$, the policy tends to avoid negative terminal states early, even if doing so requires taking longer paths, because long-term gains outweigh short-term penalties. At $\gamma = 0.5$, avoidance remains but is less extreme. At $\gamma = 0.1$, the agent may take more direct paths or tolerate risky states because only immediate rewards strongly influence decisions.