۱. ارزیابی:

- هدف اصلی الگوریتم Q-Learning چیست؟
- هر یک از هایپرپارامترهای الگوریتم Q-Learning چه کاربردی دارند؟ (آلفا، گاما و اپسیلون)
- طریقه عملکرد سیاست epsilon-greedy به چه صورت است؟
- چرا exploration در مسائل یادگیری تقویتی مهم است؟
- آیا نیاز است تا در محیط غیر قطعی و تصادفی از سیاست تصادفی استفاده کنیم؟

۲. دست گرمی:

- یک محیط grid به ابعاد 5x5 ایجاد کنید.
- در محیط ساخته شده موانعی قرار دهید تا agent نتواند از آنها عبور کند.
- به ازای هر قدم agent و برخورد آن با موانع، reward مناسب طراحی کنید.
- برای محیط قابلیت تصادفی و غیرقطعی بودن قرار دهید تا گاهی باعث حرکت تصادفی agent شوند.
- با استفاده از الگوریتم‌های بهینه‌سازی، مقادیر مناسب هایپرپارامترهای آلفا، گاما و ضریب کاهش اپسیلون را پیدا کنید.