Bytebricks Python Backend LLM Developer Live-Coding Interview

Overview

During this 45-minute session, we'll:

1. Discuss your take-home submission (10 minutes)
2. Implement one targeted enhancement (30 minutes)
3. Test and review the change (5 minutes)

AI Agent Coding Task (select the relevant one)

LangGraph Framework Tasks

Task	Description	Success Criteria
Add RAG Confidence Logging	Modify agent/rag.py to log relevance and coverage scores to console	When asking "What is RAG?", console shows: [DEBUG] RAG scores - relevance: 0.85, coverage: 0.79
Adjust Decay Factor	Change decay from 0.5 → 0.3 in adaptive.py	After /bad_answer, demonstrate score decays by 0.3 each turn via logs
Add Source Tracking	Enhance citations to include page/section numbers	Responses show: [Source: AI Basics, Page 12, Section 3.1]

OpenAI Agents SDK Tasks

Task	Description	Success Criteria
Guard-rail on Low Relevance	Add callback to block responses when relevance < 0.4	Out-of-scope query returns "not covered" Console shows: [BLOCKED] relevance=0.35
Print Token Usage	After runner.run(), display token consumption	Console shows: Tokens used: 193 (prompt: 150, completion: 43)
Add Retry Logic	Implement exponential backoff for API calls	Demonstrate retry on simulated failure with logged delays

Backend Coding Task

FastAPI/General Tasks

Task	Description	Success Criteria
Health Check Endpoint	Add /health public endpoint	GET /health returns {"status": "ok", "version": "1.0.0"}
Metrics Endpoint	Add /metrics public endpoint for basic stats	Returns request count, average response time

Enhanced Rate Limiting	Implement a global rate limit on the overall system	Apart from per minute per user limit the system restricts X requests per minute globally