

1. Which projects have you worked on?
2. Have you worked on Python?
3. Have you worked with FastAPI?
4. What is the Python and FastAPI workflow?
5. How did you select the model?
6. Where is the authentication part?
7. Are you just blindly calling the FastAPI without any authentication?
8. How are you validating the JSON output that is coming out?
9. How would you structure a Python service that supports multiple external connectors?
10. How do you handle an async API?
11. How would you implement retries?
12. If you have a system where the document size is very large, how do you handle it without loading everything into memory?
13. Suppose we have created a document integration platform. Where would you use GenAI, and where exactly would you make sure you are not using GenAI?
14. Suppose the LLM extracts all the information incorrectly. How do you handle that?
15. What is your understanding of Azure AI Foundry?
16. Have you used Azure AI Foundry and created something using Foundry?
17. How would you design the whole document integration architecture?
18. How will you receive the document?
19. How will the Python services process the document?
20. How will GenAI extract and validate information?
21. How will you take the appropriate next steps?
22. How will you send the processed information to external systems/connectors?
23. Explain prompt engineering, prompt documentation management, and configuration management.
24. In this whole architecture, which parts should be deterministic and which parts should use LLMs?