

1. What GenAI tech stack have you used?
2. Which Python frameworks have you used?
3. What is synchronous programming?
4. What is asynchronous programming?
5. Does asynchronous programming provide true parallelism?
6. What is the difference between concurrency and parallelism?
7. For CRUD operations, would you use synchronous or asynchronous programming?
8. How would you design an asynchronous operation for a complex system?
9. If 100 requests hit the same database simultaneously, how would you handle it?
10. Which Python libraries do you use for database operations?
11. Draw the architecture for a CRUD operation.
12. Explain the complete CRUD data flow.
13. How would you store and fetch data in a GenAI application?
14. Draw the GenAI/RAG architecture.
15. How do you store knowledge for a GenAI application?
16. Where do you store the original documents?
17. Where do you store embeddings?
18. What chunking strategy do you use?
19. What is chunk overlap and why is it required?
20. What is Top-K retrieval?
21. Why do you use reranking?
22. What metrics do you use to evaluate a RAG system?
23. How do you detect hallucination?
24. What is faithfulness in RAG evaluation?
25. What is the difference between context precision and context recall?
26. Which tools do you use for RAG evaluation?
27. What is KV caching?
28. What are the advantages and limitations of KV caching?
29. What is the difference between AI, ML, and GenAI?
30. What cloud platforms have you worked with?
31. How would you design a GenAI application on Azure?
32. What Azure services have you used for GenAI?
33. Why would you choose Azure OpenAI?
34. Why FastAPI for GenAI applications?
35. How would you handle 100 concurrent GenAI requests?
36. What is LangChain?
37. What is LangGraph?
38. Why would you use LangGraph instead of LangChain?

<https://www.instagram.com/rawhul213>

39. How does an AI agent work?
40. How do you implement tool calling in LangGraph?
41. How do you handle failures in an AI agent?
42. How do you reduce hallucinations in a RAG application?
43. How do you improve retrieval accuracy?
44. How do you handle documents belonging to different years or time periods?
45. How do you identify the intent of a user query programmatically?
46. How would you retrieve information for a specific year from a large document collection?
47. How do Server-Sent Events (SSE) work in a GenAI application?
48. How would you stream an LLM response to the UI?
49. How do you handle long-running GenAI requests?
50. How would you design a production-ready RAG system?

@RAWHUL213