

AI / GenAI / RAG

1. Can you explain the process of creating a RAG pipeline?
2. Suppose we have a large PDF and need to perform chunking. What is your chunking strategy?
3. Suppose you have built a RAG pipeline. How would you deploy it to production?
4. Have you deployed a RAG pipeline to production? If yes, explain the deployment process.
5. What infrastructure/cloud services have you used for deploying a RAG pipeline?
6. What changes would you make when moving a RAG pipeline from development/prototype to production?
7. How would you validate the performance of a RAG system in production?

RAG Performance / Caching

8. How would you implement caching in a RAG pipeline to improve slow response times?
9. Where exactly would you add caching in the RAG flow?
10. If every user asks the same question differently, how would caching work?
11. How would you implement semantic caching when the cache key cannot be a constant string?
12. What would happen if the vector database contains a very large dataset? How would you optimize retrieval?
13. Behind the scenes, how does a vector database perform a search using embeddings?
14. What algorithm/strategy does a vector database use for searching large numbers of vectors?
15. What is ANN (Approximate Nearest Neighbor) search?
16. What is HNSW and how does it work?
17. What is IVF / Inverted File Index?

Cloud / AI Tools

18. What cloud APIs or cloud services have you used in your projects?
19. Which cloud services have you implemented in your projects?
20. Have you used Claude Code?

21. How do you use an AI coding agent like Claude Code in your day-to-day development?
22. When you start your day, where does the AI coding agent help you in your development workflow?
23. Which model are you currently using and why did you choose that particular model?
24. Which Claude model have you used?
25. What cloud services have you explored?

Coding Agent / Token Optimization

26. If you're using an AI coding assistant, how do you efficiently manage token usage?
27. How would you reduce token consumption when using a coding agent?
28. Suppose you have an entire codebase. How can you provide only one folder's access to the AI coding assistant?
29. How can you restrict a coding agent so that it can access only a particular folder?
30. Can this be done through a setting or cloud API?
31. Do you need to create a separate repository just to give an AI coding agent access to one folder?
32. How would you provide folder-level access to a coding agent when working traditionally from Visual Studio/VS Code?
33. What is the role of project instruction files such as CLAUDE.md?
34. Can instructions in the chat itself replace maintaining a project instruction file?
35. What is the difference between restricting an agent through instructions and actually restricting filesystem access?

MCP / Production / Monitoring

36. Suppose you created an MCP and deployed it to production. How would you monitor it?
37. How would you monitor token consumption for an MCP/LLM application in production?
38. How would you monitor embedding-token usage and LLM-token usage?
39. Which monitoring tools would you use for an LLM/RAG/MCP application?

40. How would you prevent an LLM from answering questions outside the domain of your RAG application?
41. Suppose your RAG application is for company policies, but the user asks it to write a Python program. How would you prevent the LLM from answering using its general knowledge?
42. How would you handle out-of-domain questions in a RAG system?
43. How would you control token usage in production?
44. What metrics would you monitor for an LLM/RAG application?

MCP Regression / Validation

45. Suppose you have 10–12 MCP servers deployed in production and you make a change to one MCP. How would you ensure the other MCPs are still working correctly?
46. How would you validate changes when LLM responses are different/non-deterministic every time?
47. How would you verify that a change actually improved the accuracy of the response?
48. How would you perform regression testing for multiple MCP servers?
49. How would you compare a new RAG/LLM version against the previous production version?
50. What metrics would you use to determine whether the new version is better?

Payment Gateway

51. Do you have experience with payment gateway integration?
52. Have you integrated any payment gateway in your previous projects?
53. Can you explain the typical payment gateway integration flow?
54. How would you handle payment confirmation, webhooks, retries, duplicate callbacks, and transaction verification?

Python / DSA

55. How would you solve the missing-number problem?
56. What is your approach to finding a missing number?
57. Why would you use XOR for the missing-number problem?
58. If multiple numbers are missing, how would you approach the problem?

59. If the array contains 1,000,000 numbers, can you store them in memory and iterate over them?

60. What would you do if memory is a constraint when finding multiple missing numbers?

Complexity

61. What is $O(n)$?

62. What does Big-O mean?

63. What does the "O" in Big-O stand for?

64. What is the time complexity of your solution?

65. What is the space complexity of your solution?

Python enumerate()

66. What is enumerate() in Python?

67. What does enumerate() basically do?

68. In the missing-number code, what does missing represent?

69. What does this line do: $\text{missing} \wedge= i \wedge \text{num}$?

70. Can you explain the dry run of the XOR solution?

71. How does XOR help us find the missing number?

This interview has mainly covered RAG architecture, production deployment, vector databases, caching, LLM observability, MCP, AI coding agents, cloud services, Python, and DSA.