

AI / Production System

1. What is the most complex production AI system you have worked on?
2. Can you explain the architecture of that production AI system?
3. Is the system binary/deterministic or random?
4. Can the incoming input be video or images?
5. How do you handle intelligence/unknown or arbitrary user input?
6. How do you handle policy errors?
7. What is a parsing error?

RAG

8. What is RAG?
9. Explain the RAG architecture.
10. Draw the RAG architecture.
11. What happens in the RAG retrieval pipeline?
12. What are embeddings?
13. What is semantic meaning?
14. What is semantic search?
15. What is Top-K retrieval?
16. Why do we use reranking?
17. What is reranking?
18. What is the output of reranking?
19. How did you use reranking in your implementation?
20. How do you calculate relevance/similarity score?
21. What happens if retrieved chunks are duplicates?
22. How do you perform deduplication in RAG?

PDF / Document Processing

23. How do you handle complex PDF layouts such as tables, headers, footers, and multi-column text?
24. Have you used any preprocessing libraries for PDFs?
25. Which PDF libraries have you used?
26. How do you handle scanned PDFs?
27. How do you handle table extraction from PDFs?

Chunking

28. How do you chunk documents?
29. Do you chunk paragraph-by-paragraph or sentence-by-sentence?
30. What chunking strategies have you used?
31. How do you avoid splitting in the middle of a sentence?
32. Have you faced challenges while choosing a chunking strategy?
33. What is semantic chunking?

34. Why do you use chunk overlap?

35. What factors do you consider when choosing chunk size?

LangChain

36. What is LangChain?

37. What are the major components of LangChain?

38. How do you use LangChain in a RAG pipeline?

39. Which LangChain text splitter have you used?

Python / Programming

40. Write a deduplication function.

41. Write a deduplication function for retrieved RAG chunks.

42. What is the time complexity of the deduplication function?

43. What is the space complexity?

44. Is space complexity related to memory?

45. Why is the space complexity $O(n)$?

46. What is the difference between time complexity and space complexity?

47. Which Python library do you use for deduplication?

Libraries / Technology Stack

48. Which library do you use for PDF processing?

49. Which library do you use for chunking?

50. Which library do you use for deduplication?

51. Which library do you use for relevance/similarity scoring?

52. Which library do you use for reranking?

53. Which embedding libraries/models have you used?

54. Which vector databases have you used?

Important follow-up questions they can ask

55. Why did you choose Top-K?

56. Why do you need reranking if vector search already gives relevance scores?

57. What is the difference between similarity score and reranking score?

58. How do you improve retrieval accuracy?

59. How do you reduce hallucinations in RAG?

60. How do you handle irrelevant queries?

61. How do you handle prompt injection?

62. How do you evaluate a RAG system?

63. What metrics do you use for retrieval quality?

64. How do you monitor a production RAG system?