Assignment 16
System Design & Scenario Questions

Part D — Interview Scenarios

16 questions

★ **Must-know** — asked very frequently in GenAI / ML interviews (7 questions). Prepare these first.

1. Design a chatbot that answers questions over 10,000 internal company PDFs. Cover ingestion, chunking, embeddings, vector DB, retrieval, reranking, LLM choice, evaluation and monitoring. ★ · Asked: Deloitte, EY, Google
2. Your RAG chatbot answers correctly in testing but users complain in production. How do you investigate? ★ · Asked: Google
3. Design a customer-support agent that can look up orders and issue refunds safely (tools, permissions, human approval). · Asked: Google
4. Your LLM API bill has tripled this month. What would you check and change? · Asked: EY
5. A user's question needs data from both a SQL database and documents. Design the system (Text-to-SQL + RAG routing). · Asked: Deloitte
6. How would you build a multilingual (e.g. Hindi/Marathi + English) GenAI assistant?
7. Design a document-extraction pipeline that pulls structured fields from invoices/forms with an LLM.
8. How would you handle confidential data when using a third-party LLM API? · Asked: Deloitte, Google
9. P95 latency of your assistant is 12 seconds. How do you bring it under 3 seconds? · Asked: TCS, Google
10. How would you A/B test two prompts or two models in production? · Asked: Google
11. Your production RAG chatbot is confidently hallucinating information that does not exist in the knowledge base, and the client is escalating. How do you diagnose and fix it? ★ · Asked: EY
12. The knowledge base grew from 10,000 to 5 million documents. Retrieval latency crashed and hosting cost spiked. What do you change? ★ · Asked: EY
13. A non-technical client insists the chatbot must be "100% accurate and never make a mistake". How do you handle this without jargon? ★ · Asked: EY, Deloitte, Google
14. Production responses suddenly became worse right after a deployment. What do you investigate? ★ · Asked: Google

15. A customer wants a multi-agent architecture "because it sounds advanced". What do you recommend? · Asked: Google

16. Design a system that processes and summarizes 10 million documents. ★ · Asked: Google