Assignment 11
RAG & Vector Databases

Part C — Building GenAI Applications

30 questions

★	**Must-know** — asked very frequently in GenAI / ML interviews (17 questions). Prepare these first.
[CODE]	**Hands-on task** — write and run the code (3 tasks).

1. What is RAG? Explain its architecture (indexing pipeline + retrieval + generation) with a diagram. ★ · Asked: TCS, PwC, Infosys, Google
2. RAG vs fine-tuning vs long-context prompting — when would you use each? ★ · Asked: TCS, Infosys, Google
3. What are embeddings? Name OpenAI embedding models and open-source embedding models. Can you reduce OpenAI embedding dimensions? ★ · Asked: Google
4. How do you choose an embedding model (MTEB leaderboard, domain, language, dimensions, cost)?
5. What are chunking strategies (fixed, recursive, semantic, document-structure based)? How do you choose chunk size? ★ · Asked: TCS, Deloitte, Google
6. How are indexes created in vector databases? Explain flat, IVF, PQ and HNSW. ★
7. How does HNSW (Hierarchical Navigable Small Worlds) work? ★
8. Compare vector databases: Chroma, FAISS, Pinecone, Qdrant, Weaviate, Milvus, pgvector. · Asked: TCS, Google
9. Store embeddings in ChromaDB, persist them to disk and load them back. [CODE]
10. Create retrievers that fetch: (a) top-5 records, (b) records with similarity score > 0.5, (c) results using MMR. [CODE]
11. What is MMR (Maximal Marginal Relevance) and why use it? ★ · Asked: Wipro
12. What is hybrid search (BM25 + vector)? Why does it help? ★ · Asked: Google
13. What is re-ranking? Explain cross-encoders and Cohere Rerank. ★ · Asked: Wipro, Google
14. Explain MultiQueryRetriever, ContextualCompressionRetriever, ParentDocumentRetriever and SelfQueryRetriever. · Asked: Google
15. What are query transformation techniques (HyDE, query rewriting, query decomposition)? · Asked: Wipro, Google
16. Build a complete retrieval chain (retriever + prompt + LLM) that cites its sources. [CODE]
17. How do you build RAG using an open-source LLM end-to-end?

18. How do you feed table data and image data to a RAG system? · Asked: Google
19. What is Agentic RAG? What is GraphRAG? ★ · Asked: Google
20. Your RAG system gives wrong answers. How do you debug it step by step (retrieval vs generation)? ★ · Asked: Deloitte, Wipro, Google
21. Explain some real-world RAG use cases.
22. If frontier LLMs have huge context windows, why not send the entire PDF (or all documents) to the model instead of building RAG? ★ · Asked: Deloitte, Google
23. What is the difference between tokenization and chunking? ★ · Asked: PwC
24. Can a GenAI system work without a vector database? When is a vector DB not needed? · Asked: PwC
25. How do you handle restricted vs unrestricted documents in RAG, so unauthorized documents never enter the LLM context? ★ · Asked: TCS, EY, Google
26. How do you implement citations in RAG? And what if the answer is correct but the cited source is wrong? ★ · Asked: Google
27. What happens if no relevant document is retrieved? How do you reduce hallucination when retrieval is weak? ★ · Asked: Google
28. What is grounding / grounded generation? · Asked: Google
29. What is RAG poisoning (knowledge-base poisoning), and how do you prevent it? · Asked: Google
30. Walk through everything that happens behind the scenes when a user submits a query to your RAG app, and where latency or failures can occur. ★ · Asked: Deloitte, TCS