1. What are hallucinations in LLMs, why do they occur, and how do you detect and mitigate them in production systems?

2. In a high-throughput production system, when would you choose a list over a dictionary (or vice versa), and why?

3. In a performance-critical Python service, how do best- and worst-case list operations impact latency?

4. Which similarity measures are used to compare embedding vectors, and how do you choose the right one for production RAG systems?

5. How are embedding vectors stored and indexed inside a vector database, and what trade-offs affect speed, memory, and recall?

6. What is re-ranking in a RAG pipeline, when should it be introduced, and what trade-offs does it create?

7. How would you design a secure enterprise chatbot with role / pay-grade based access control?

8. How is vector retrieval fundamentally different from SQL queries?

9. What is LangGraph, how does it differ from LangChain, and how does agentic RAG change system design?

10. In a production RAG system, do you perform similarity search immediately, or apply validation before retrieval?

11. How do you guarantee valid JSON output from an LLM in production?

12. How do you improve reasoning quality without exposing chain-of-thought?

13. When would you choose FAISS over a managed vector DB like Pinecone or Weaviate?

14. Explain inverted indexes and product quantization in FAISS and justify your index choice.

15. With ~800K embeddings, what compute-efficient retrieval strategies would you use?

16. How does a vector database retrieve relevant vectors for a query embedding?

17. How is indexing done in FAISS?

18. What are the different categories of vector stores and when should each be used?

19. When should you use RAG vs fine-tuning in production?

20. Why is FastAPI preferred for ML / LLM inference services?

21. What are the limitations of few-shot prompting?

22. What is the difference between zero-shot and few-shot prompting, and how do you decide which is better?

23. What is GPT-4’s token limit and why does it matter?

24. Which embedding model did you use and how did you evaluate the choice?

25. What chunking strategies are used in RAG systems and how do you choose the right one?

26. Why is chunking still required for ~500K rows even if LLMs can handle small datasets?

27. How would you describe the size and structure of your dataset and why did it matter?

Hallucinations

1. RAG chatbot returns confident but incorrect answers — how do you debug without increasing latency?

2. In healthcare/finance, how do you prevent hallucinations from reaching users?

Python Lists / Dicts

3. RAG pipeline latency spikes due to list usage — which operations cause worst-case behavior?

4. Memory pressure from dicts at scale — when would you switch to a different structure?

Similarity Measures

5. Switching embedding models drops retrieval quality — cosine still correct?

6. Multimodal embeddings have different vector norms — how do you compare safely?

Vector Storage

7. Vector DB grows from 1M → 100M vectors — how do you control memory cost?

8. Exact search violates latency SLOs — redesign strategy?

Re-Ranking

9. Re-ranking doubles latency — how do you rebalance accuracy vs speed?

10. Re-ranking is expensive — when do you skip it dynamically?

Access Control RAG

11. Junior employee sees restricted data — where did the system fail?

12. Roles change monthly — how do you update access without re-embedding?

SQL vs Vector Search

13. Replace vector DB with Postgres + pgvector — where does it break?

14. SQL filters reduce recall — how do you redesign retrieval order?

LangGraph / Agentic RAG

15. Why LangChain agents fail for loops and approvals?

16. Agent over-retrieves and increases cost — how do you control it?

Retrieval Order

17. System retrieves for “hi” messages — how do you avoid waste?

18. Pre-retrieval classifier skips needed retrieval — how do you fix it?

JSON Output

19. JSON is valid but semantically wrong — how do you detect?

20. Retries increase latency — how do you enforce structure in one pass?

Reasoning Quality

21. Accuracy drops without chain-of-thought — how do you redesign safely?

22. Agent makes wrong decisions — how do you debug without inspecting thoughts?

FAISS vs Managed DB

23. FAISS can’t handle traffic growth — migration strategy?

24. Managed DB bill spikes — optimize before switching back?

FAISS Indexing

25. IVF+PQ drops recall — which parameters do you tune?

26. Frequent inserts/deletes — why is IVF+PQ problematic?

Scaling Retrieval

27. P95 latency exceeds SLO — which ANN index and knobs first?

28. ANN helps speed but raises cost — how do you reduce compute?

Vector Retrieval

29. Fast results but missing relevant docs — where do you debug?

30. Unauthorized docs must never appear — where to enforce access?

Vector Stores

31. When to migrate from FAISS to managed DB?

32. Managed DB too expensive — how do you self-host safely?

RAG vs Fine-Tuning

33. Fine-tuned policies become outdated — what went wrong?

34. RAG breaks latency SLOs — what do you replace with fine-tuning?

FastAPI

35. High concurrency + slow inference — how do you avoid blocking?

36. API schema changes — how do you handle backward compatibility?

Prompting

37. Few-shot works in demo but fails in prod — what's your decision framework?

38. Few-shot increases cost 35% — alternatives?

Token Limits

39. Context overflow in RAG — architectural fix?

40. Larger context improves accuracy but doubles cost — redesign?

Embeddings

41. Switching embedding model breaks retrieval — debugging steps?

42. Better embeddings double latency — how do you decide?

Chunking

43. Over-chunking loses context — how do you fix it?

44. PDFs with tables + images — modality-aware chunking strategy?

Dataset Scale

45. Same row count, worse retrieval — why?

46. Schema evolves — how do you update without full re-index?

Ingestion / Azure Flow

1. **How does the system detect and ingest PDF attachments from Outlook using Microsoft Graph API?**

2. **Why is Azure Blob Storage used in the document ingestion pipeline? What problem does it solve?**

3. **What is Microsoft Graph API and how is it used to access Outlook emails programmatically?**

4. **What is the role of Azure Functions or Azure App Service in this architecture, and how do you decide between them?**

OCR / Document Processing

5. **How does OCR work in production systems, and which LLM-based OCR solutions have you used?**

6. **After extracting data from PDFs, what downstream processing steps does the system perform before making it searchable?**

7. **What is document parsing, and how do you handle parsing across heterogeneous document types and data sources?**

RAG / Multimodal / Search

8. **How would you design a multimodal RAG system (text, images, tables), and what challenges arise compared to text-only RAG?**

9. **What is Azure AI Search, and how does it fit into a RAG architecture?**

10. **What is Azure AI Studio, and how is it used in building or managing GenAI applications?**

11. **What is RAG architecture, and how does data flow through retrieval, generation, and validation stages?**

12. **How does RAG architecture work end-to-end in a production system?**

Vector DB / Embeddings / Retrieval

13. **Which vector database did you use, and what factors influenced that choice?**

14. **How do you design and manage metadata in a vector database for filtering, security, and relevance?**

15. **What is a vector database, and why is it required for semantic search or RAG systems?**

16. **What are embeddings, and which embedding model did you use? Why was it suitable for your use case?**

17. **How is the retrieval operation performed in a vector-based RAG system?**

Databases / State

18. **Why do we use Azure Cosmos DB in GenAI systems, and what type of data is best suited for it?**

19. **How do you design state management and caching in an AI system?**

20. **How do you handle memory management in conversational AI systems?**

LLMs / Models / Comparison

21. **What is the difference between BERT and large language models (LLMs) in terms of architecture and use cases?**

22. **How is an LLM fundamentally different from a BERT-based model?**

23. **How is an LLM-based chatbot different from a traditional rule-based or intent-based chatbot?**

24. **Which LLMs have you worked with, and why did you choose them for specific projects?**

25. **Have you worked with Google Gemini models, and in what scenarios did you evaluate or use them?**

26. **How is Gemini different from GPT-4 in terms of architecture, capabilities, and use cases?**

27. **What was the typical token size of inputs you sent to the LLM, and how did it influence system design?**

Deployment / Cloud

28. **How would you deploy a Gemini-based RAG application on GCP and Azure? What services would you use?**

29. **How do you manage concurrency when many users interact with an AI system simultaneously?**

Agents / LangChain / LangGraph

30. **What is LangGraph, and how does it differ from LangChain?**

31. **What is the difference between RAG and agent-based systems, and when do you combine them?**

32. **How do you design an agentic flow for complex decision-making tasks?**

33. **What applications have you built using RAG, LangChain, and LangGraph?**

Safety / Governance

34. **How do you define and enforce guardrails for AI responses in production?**

35. **How do you implement explainable AI and responsible AI practices in real-world AI systems?**

Evaluation / Accuracy

36. **How do you evaluate model accuracy and system-level performance for LLM-based applications?**

Conversational / Voice AI

37. **Have you worked on conversational chatbots or voice bots, and what architectural differences did you encounter?**

System Design / Experience

38. **How would you design a full-fledged conversational AI system end-to-end?**

39. **Can you explain the key AI use cases you've worked on, including the business problem and technical solution?**

Backend & API (Very Common in Fidelity)

1. **Why did you choose FastAPI for building your backend services? What advantages did it give you in production?**

2. **How is a REST API designed for scalability and versioning in enterprise systems?**

3. **How did you use SQLAlchemy in your project, and why not raw SQL everywhere?**

4. **Explain SQL concepts from basic to intermediate level that you used in your project (joins, indexes, transactions).**

🔹 Retrieval / RAG (Core Enterprise GenAI Topic)

5. **Explain the retrieval process in your system end-to-end.**

6. **How does your retrieval pipeline decide what data to fetch before sending it to the LLM?**

7. **How would you explain the retrieval process to a non-technical stakeholder?**

✅ (Yes bhai — ye question REAL hai, business communication test karta hai)

🔹 Use Case Discussion (Lilly / Fidelity Style)

8. **Explain the business problem you solved in the Lilly/Fidelity project. Why was AI required?**

9. **What was your technical approach to solving this business problem?**

10. **What alternatives did you consider, and why did you choose this approach?**

🔹 Validation & Quality Checks (Very Important)

11. **How did you validate that the questions or user queries were correct before processing them?**

12. **How did you ensure data quality and correctness in the retrieval results?**

🔹 Agents / Agentic Flow (Advanced but Valid)

13. **What agents were involved in your system, and what responsibility did each agent have?**

14. **How do agents communicate and coordinate in your architecture?**

🔹 System Explanation Skills (Enterprise MUST)

15. **Explain the entire system flow to a non-technical audience (product manager / business stakeholder).**

Introduction & Project Context

1. **Can you briefly introduce yourself and walk us through the AI/ML project you are currently working on?**

2. **What business problem does your current Generative AI project solve, and what is your role in it?**

3. **How have you used LangChain and LangGraph in a real production workflow?**

🔹 Traditional Machine Learning Experience

4. **Have you worked on traditional machine learning models in industry-grade projects? If not, how have you applied them in POCs or controlled environments?**

5. **How comfortable are you transitioning between traditional ML approaches and GenAI-based solutions?**

🔹 Linear Regression & Statistics

6. **What are the core assumptions of linear regression, and why are they important?**

7. **If residuals show a fan-shaped pattern, what assumption is being violated and how would you statistically verify it?**

8. **Which statistical tests are used to detect heteroscedasticity in regression models?**

🔹 Logistic Regression & Metrics

9. **What evaluation metrics are commonly used for logistic regression, and why?**

10. **In what scenarios would you prioritize precision over recall, or recall over precision?**

11. **Does precision remain constant across different probability thresholds? Why or why not?**

🔹 GenAI Project Deep Dive

12. **Can you walk us through your GenAI project end-to-end, including data ingestion, retrieval, and response generation?**

13. **What specific components did you personally design or implement in this project?**

14. **How did you collaborate with other teams or roles while owning multiple parts of the system?**

🔹 Frameworks & Tools

15. **Which AI frameworks and tools did you use in your project, and what were the reasons behind those choices?**

16. **Why did you choose LangGraph over a purely chain-based or prompt-based approach?**

🔹 RAG-Based Chatbot Design

17. **How would you design a chatbot using Retrieval-Augmented Generation (RAG) from scratch?**

18. **What factors influence your chunking, embedding, and retrieval strategy in a RAG system?**

19. **Which document parsing libraries have you used, and how do you decide between them?**

20. **How do you design prompt templates to ensure consistent and structured outputs?**

21. **How do you enforce structured outputs such as JSON in LLM responses?**

🔹 Validation & Quality Control

22. **How do you validate user queries before sending them through the retrieval and generation pipeline?**

23. **How do you verify the correctness and relevance of retrieved documents?**

🔹 Agents & Agentic Workflows

24. **What agents were involved in your system, and what responsibilities did each agent have?**

25. **How do agents coordinate and share state in an agentic architecture?**

🔹 Retrieval Process (Core Focus)

26. **Can you explain the retrieval process step by step in your system?**

27. **How would you explain the retrieval process to a non-technical stakeholder or business user?**

28. **What happens if the retrieval layer fails or returns irrelevant data?**

🔹 Backend & API Design

29. **Why did you choose FastAPI for building backend services?**
30. **How do you design REST APIs for scalability and maintainability?**
31. **How have you used SQLAlchemy in your project, and what advantages does it provide?**
32. **Which SQL concepts (joins, indexes, transactions) did you actively use in your implementation?**

🔹 System Design & Architecture

33. **How would you design a full-fledged conversational AI system end-to-end?**
34. **How do you manage state, caching, and session memory in AI-driven applications?**
35. **How do you handle concurrent users accessing the system simultaneously?**

🔹 Evaluation & Governance

36. **How do you evaluate the accuracy and performance of AI/ML models in production?**
37. **What mechanisms do you use to define guardrails and prevent unsafe AI responses?**
38. **How do you approach explainable AI and responsible AI in real-world applications?**

🔹 Behavioral & Motivation

39. **Why are you interested in working with our organization?**

40. **Are you open to working on optimization or non-GenAI problem statements if required?**

41. **What kind of AI/ML roles and problems excite you the most?**

Project & Ownership

1. **Can you explain your project end-to-end, including the business problem it was solving and your role in the solution?**

2. **Which parts of the project did you personally own, and where did you collaborate with other teams?**

🔹 Agents & Agentic Architecture (Very Important)

3. **What are AI agents, and why did you choose an agent-based architecture for your workflow?**

4. **How many agents were involved in your system, and what responsibility did each agent have?**

5. **How did these agents interact and coordinate with each other during execution?**

6. **What was the role of the supervisor agent, and how did it orchestrate or control the other agents?**

7. **At a code level, how were these agents implemented (classes, functions, services)?**

🔹 RAG & Architecture

8. **Can you explain the RAG architecture you implemented in your project?**

9. **How did data flow from ingestion to retrieval to response generation in your system?**

🔹 Scalability, UI & Concurrency

10. **How did your system handle multiple concurrent users accessing it at the same time?**

11. **What did the user interface look like, and how did users interact with the system?**

🔹 Challenges & Trade-offs

12. **What were the major technical or architectural challenges you faced during the project?**

13. **How did you resolve those challenges, and what trade-offs did you make?**

🔹 Document Generation Decision (Excellent KPMG Question)

14. **Why did you choose to generate Word documents programmatically instead of using the LLM’s document-generation capability directly?**

15. **What limitations or risks did you see in relying entirely on LLM-generated Word documents?**

🔹 Cloud, Model & Benchmarking Decisions

16. **Why was the LLM hosted on Azure while other parts of the stack were deployed on AWS?**

17. **Why did you choose OpenAI’s LLMs for this project?**

18. **What benchmarking or evaluation did you perform to conclude that GPT-4o was the right model?**

🔹 Career & Skill Profile

19. **Which areas of your career have you worked in—Generative AI, backend development, data analysis, or others?**

20. **How would you rate your proficiency in Python, and what kinds of problems have you solved using it?**

21. **Have you worked on the data side of Python, such as ETL pipelines, data frames, or data processing workflows?**

🔹 Conceptual AI Understanding (Business-Friendly)

22. **How does an LLM summarize a piece of text? What is happening conceptually under the hood?**

23. **Why do you think the Generative AI revolution happened in the last 2–3 years and not a decade ago?**

24. **What technical breakthroughs or limitations were responsible for enabling this shift?**

🔹 Vectors & Embeddings

25. **What are vectors and embeddings, and what do they represent in an AI system?**

26. **Why are embeddings critical for semantic search and RAG-based systems?**

🔹 Domain & Business Fit

27. **Do you have experience or familiarity with the tax domain? Would you be open to working on tax-focused AI projects?**

🔹 Logic / Thinking Test (Classic Consulting Signal)

28. **You have two doors—one leads to freedom and the other to death. Each door has a guard; one always tells the truth and the other always lies. You can ask one question to one guard. What would you ask to determine the correct door?**

Project & Architecture (Deep Technical)

1. **Can you explain your project from a purely technical perspective, focusing on architecture and design decisions?**

2. **Can you walk me through the RAG architecture you implemented end-to-end?**

🔹 Document Parsing & Ingestion (Cost-Aware Thinking)

3. **How did you parse the documents in your system, and what challenges did you face with real-world document structures?**

4. **If a table spans across multiple pages in a document, how would you parse it accurately in a cost-effective way without overusing LLMs?**

(Yes bhai — this is a classic product-company cost-awareness trap)

🔹 Embeddings & Vector Databases

5. **Which embedding model and vector database did you use, and why were they suitable for your use case?**

6. **What was the dimensionality of the embedding vectors you used, and how does vector length impact retrieval accuracy, latency, and storage cost?**

🔹 Determinism & Reliability (Very Important)

7. **How do you ensure deterministic or near-deterministic responses from LLMs in systems with strict business rules?**

8. **What techniques can be used to reduce randomness and enforce consistency in LLM outputs?**

🔹 RAG Failure Handling & Evaluation (Senior-Level)

9. **What kinds of failures can occur in a RAG pipeline, and how do you detect them?**

10. **How do you evaluate the performance of a RAG system, and where does the evaluation component sit in the overall architecture?**

11. **What strategies would you apply to continuously improve RAG accuracy and robustness?**

🔹 Similarity Search & Distance Metrics

12. **What types of similarity search can be used in a RAG pipeline, and how do you decide which one to use?**

13. **Why is cosine similarity commonly used for embeddings, and what would change if you used Euclidean or Manhattan distance instead?**

🔹 Traditional ML Fundamentals

14. **Can you explain a traditional machine learning project you’ve worked on from problem formulation to evaluation?**

15. **What are the key evaluation metrics for classification problems, and when would you prioritize one over another?**

16. **If you observe a drop in model accuracy in production, how would you investigate and respond?**

17. **How do you retrain a model, and how do you split data for training, validation, and testing when tuning hyperparameters?**

🔹 Coding (Hands-On Signal)

18. **Write a Python function to generate all possible subsets of a given list of strings.**

GSK – VERIFIED & REPHRASED QUESTION SET

(Clean, professional, global-enterprise wording)

🔹 Ingestion & Storage

Outlook PDF Detection

1. **How does your solution detect and extract PDF attachments from Outlook emails?**

Azure Blob Storage

2. **What role does Azure Blob Storage play in the ingestion pipeline, and why is it preferred over alternatives?**

Microsoft Graph API

3. **How do you use Microsoft Graph API to programmatically access and monitor Outlook mailboxes?**

Azure Functions / App Service

4. **Why did you choose Azure Functions or Azure App Service, and what responsibilities do they handle in the workflow?**

🔹 Document Understanding

OCR & LLM-Powered OCR

5. **Can you explain how OCR works conceptually, and name some OCR solutions enhanced by LLMs?**

Post-Extraction Processing

6. **After extracting content from PDFs, what downstream processing steps does your system perform before retrieval or analytics?**

🔹 Retrieval-Augmented Generation (RAG)

Multi-Modal RAG

7. **How would you design a multi-modal RAG system that handles text, tables, and images?**

🔹 Azure AI Services

Azure AI Search

8. **What is Azure AI Search, and where does it fit in your overall architecture?**

Azure AI Studio

9. **What capabilities does Azure AI Studio provide, and how would you use them in an enterprise GenAI project?**

🔹 Vector Data & Persistence

Vector Database Choice

10. **Which vector database did you choose, and what criteria influenced that decision?**

Metadata Management

11. **How do you store and leverage metadata alongside vectors to improve retrieval accuracy and governance?**

Azure Cosmos DB

12. **Why is Azure Cosmos DB part of the architecture, and what type of data is best suited for it?**

🔹 OCR Models & Trade-offs (High-Depth)

Donut vs TrOCR

13. **In which scenarios would an OCR-free model like Donut outperform an OCR-first pipeline such as TrOCR? What trade-offs should be considered (accuracy, latency,**

language coverage, model size)?

Decoder Choice in TrOCR

14. **Why does TrOCR initialize its text decoder from RoBERTa instead of GPT-2, and how does that affect training stability and fine-tuning?**

Handwriting Recognition

15. **Which architectural components make TrOCR strong at handwriting recognition, and how would you benchmark it against Tesseract + Seq2Seq?**

🔹 Layout-Aware Models

Layout Awareness

16. **How does LayoutLMv3 encode spatial relationships between tokens, and why can’t Donut or vanilla BERT achieve this without bounding-box embeddings?**

External OCR Dependency

17. **LayoutLM relies on externally generated text and bounding boxes. Describe two scenarios where this dependency is a limitation and two where it is beneficial.**

🔹 Key-Value Extraction

18. **Donut supports end-to-end key-value extraction. What training objective enables this, and how would you adapt it for invoice line-item extraction?**

🔹 Fine-Tuning Strategy

19. **Given 10K manually labeled receipts, which model (Donut, TrOCR, or LayoutLMv3) would you fine-tune and why? Explain the preprocessing pipeline, loss functions, and expected challenges.**

🔹 Compute & Cost

20. **Compare the GPU/CPU requirements of Donut and LayoutLMv3 for large-scale inference (e.g., 100K documents/hour). How would you optimize each?**

🔹 Hybrid Pipelines

21. **Design a hybrid pipeline combining Donut for coarse key-value extraction and LayoutLMv3 for fine-grained validation. How does data flow between the models?**

🔹 Transfer Learning

22. **How would you perform cross-language transfer for Donut on low-resource scripts such as Devanagari? Discuss synthetic data generation, tokenization, and evaluation.**

🔹 Model Evaluation

23. **Beyond character-level accuracy, what composite metrics would you use to evaluate the real-world effectiveness of document understanding models?**

BCN Associate – Org COE L1

🔹 Motivation & Career Decisions (High Pressure Area)

1. **Why Bain? What specifically attracts you to Bain Capability Network?**

2. **Why are you leaving PwC, and what gaps are you hoping to address at Bain?**

3. **What aspects of your role at PwC did you enjoy, and what aspects did you find limiting?**

(Expect multiple uncomfortable follow-ups — yes, this is intentional)

🔹 Guesstimates (Classic Consulting Signal)

4. **Estimate the monthly revenue of a salon in Bangalore. Walk us through your assumptions.**

🔹 Case Interview

5. **A US-based manufacturing firm wants to shift its operations to India. What factors should it consider before making this decision?**

6. **How would your recommendation change based on cost, talent availability, regulatory environment, and supply chain risks?**

🔹 Experience & Domain Understanding

7. **Can you explain your most recent project in detail?**

8. **Walk us through your career trajectory at PwC and how it prepared you for this role.**

9. **Do you have an understanding of the organization or operating model domain? How have you applied it in your work?**

BCN Associate – Org COE L2

🔹 Project Depth (More Ownership Expected)

10. **Can you explain the project you worked on, focusing on problem definition, approach, and outcomes?**

🔹 GenAI / Analytics Maturity

11. **How did you identify and handle hallucinations in your AI or analytics solution?**

12. **Which models, tools, or frameworks did you use, and why were they appropriate for the problem?**

🔹 Responsible AI & Governance (Very Bain-Specific)

13. **How did you implement human-in-the-loop mechanisms and ensure accountability in your solution?**

14. **How did you incorporate responsible AI principles, especially in scientifically sensitive or high-impact contexts?**

🔹 Advanced Guesstimates

15. **Estimate the annual revenue of Netflix. Explain your approach and assumptions.**

🔹 Bain Motivation (Deep Grilling)

16. **Why Bain? Why not continue at PwC?**

17. **How did you first learn about Bain, and what aspects of the firm resonate with you?**

18. **What differentiates Bain from other consulting firms in your view?**

🔹 Self-Awareness & Feedback (Critical Bain Signal)

19. **If we spoke to your manager today, what three strengths would they highlight about you?**

20. **What three areas would they suggest you improve on, and why?**

🔹 Team & Collaboration

21. **What was the structure of your project team, and where did you fit within that structure?**

22. **How did you interact with stakeholders across different levels of the organization?**

🔹 Project & Architecture

1. **Can you explain the overall architecture of the project you are currently working on?**

🔹 RAG + LangGraph System Design (Core Question)

2. **Design a RAG-based chatbot using LangGraph where documents are hosted in an external data service and are continuously added, updated, or deleted. How would**

you handle ingestion, retrieval, and hallucination control?

3. **How would your LangGraph workflow handle query analysis, query rewriting, document relevance checks, and guardrails?**

🔹 Framework Choices

4. **Have you worked with CrewAI? Why did you choose LangGraph for your project instead of other agent frameworks?**

5. **What alternative frameworks exist for building agentic systems, and how would you approach a project if a client asked you to use a framework you haven't worked with before?**

🔹 LLM Selection

6. **Which LLM did you use in your project, and what factors influenced that choice?**

🔹 Traditional Machine Learning (Deep Dive)

7. **Can you explain one of your ML projects in detail, focusing on the model you used and why?**

8. **You mentioned using Random Forest. Why was it suitable for your problem, and how did you tune it?**

9. **How do you handle class imbalance in classification problems?**

10. **Can you explain precision and recall, and how would you communicate these concepts to a non-technical business stakeholder?**

🔹 Deep Learning Fundamentals

11. **Which optimizers have you worked with in deep learning, and how do you decide which one to use?**

12. **Can you explain how the Adam optimizer works conceptually?**

(Yes bhai — very common even at L1)

🔹 Python Fundamentals & Performance

13. **Are you familiar with dynamic programming? Where would you apply it in real problems?**

14. **When would you use tuples instead of lists in Python?**

15. **Explain the difference between the `in` operator and the `==` operator. How do their performance characteristics differ?**

🔹 Python Coding (Hands-On)

16. **Given a dictionary where values are tuples, perform operations such as filtering or aggregation on the data. Please share your screen and code the solution.**

17. **Design a simple banking system in Python that supports user onboarding, deposits, withdrawals, lending, and balance display. Focus on clean design rather than UI.**

*(Yes — this is a classic **LLD + Python** test)*

🔹 NLP vs GenAI Decision-Making

18. **Have you worked with NLP tasks such as NER or text classification?**

19. **Can you describe a scenario where you would recommend a traditional NLP or deep learning solution instead of a generative AI approach? Why?**

KPMG GCC – L1

VERIFIED & REPHRASED INTERVIEW QUESTIONS

🔹 Cloud Stack & Azure Expertise (Entry Gate)

1. **Which cloud stack are you currently working on, and what Azure services have you used in GenAI projects?**

2. **How familiar are you with Azure services such as Azure AI Search, Document Intelligence, Azure AI Studio, and Azure OpenAI?**

🔹 RAG Chatbot – Full System Design

3. **You are asked to design a chatbot that answers questions from hundreds of proprietary documents. Walk me through the complete technical design end-to-end.**

4. **How would you approach document analysis and ingestion, including parsing, chunking, indexing, and metadata design?**

5. **How would you design the retrieval pipeline—vector database choice, indexing strategy, filtering, and search approach?**

6. **How would you orchestrate the overall workflow using LangChain or LangGraph, including query analysis and guardrails?**

🔹 Memory Management & Conversational Context

7. **How would you manage conversational memory in your chatbot to support follow-up questions within a session?**

8. **What storage options would you consider for conversation memory, and what trade-offs exist between Redis, Postgres, or other databases?**

(Yes bhai — interviewer intentionally challenged your Redis choice)

🔹 Performance & Latency Troubleshooting

9. **If users report performance issues with the chatbot, how would you diagnose the problem?**

10. **Assuming latency is the issue, what techniques would you use to reduce response time at scale?**

🔹 Python & SQL Fundamentals

11. **How would you rate your proficiency in Python and SQL, and why did you choose that rating?**

12. **Can you explain different types of SQL joins and when to use them?**

13. **How would you identify duplicate records in SQL?**

14. **Can you explain a self-join and a real-world use case for it?**

🔹 BI & Learning Ability

15. **You’ve mentioned Power BI on your resume. How did you decide to use it, and how did you learn it effectively?**

16. **How do you generally approach learning a new tool or technology in a project environment?**

🔹 ML Project Architecture

17. **You mentioned an ML project—can you explain its architecture from data sources to model deployment?**

18. **How did you orchestrate data cleaning, transformation, and feature engineering?**

19. **Where did you store intermediate datasets, and why?**

20. **Which models did you use, and how did you evaluate them?**

🔹 APIs & Ownership

21. **Which modules of the project did you personally develop?**

22. **Did you work on backend APIs, and what responsibilities did those APIs handle?**

🔹 Vector Databases

23. **Which vector databases have you worked with?**

24. **Have you used ChromaDB or Pinecone, and in what scenarios would you prefer one over the other?**

🔹 Cross-Cloud Design Decisions (Very KPMG-Specific)

25. **You mentioned using Azure OpenAI while the rest of the stack was on AWS. Why was this decision made?**

26. **Why didn’t you use LLMs hosted directly on AWS?**

27. **How did you integrate Azure services into an AWS-based application, and what challenges did you face?**

🔹 Candidate-Led Discussion (Strong Move)

28. **Do you have any questions for us?**

(Using this to discuss interviewer’s project and brainstorm solutions = ***very strong signal****, by the way)*