GenAI Interview Must-Know Questions

The 145 questions that come up in almost every GenAI / Agentic AI interview

Drawn from real interviews at TCS, Infosys, Wipro, Deloitte, EY, PwC, EPAM, Google and product companies

★ **Must-know** — asked very frequently in GenAI / ML interviews (145 questions).

Contents

PART A — FOUNDATIONS	
1. NLP Basics & Text Preprocessing	8 Qs
2. Text Representation & Word Embeddings	10 Qs
3. Neural Network Fundamentals	8 Qs
4. Sequence Models: RNN, LSTM, GRU	6 Qs
PART B — TRANSFORMERS & LLMs	
5. Attention & Transformer Architecture	9 Qs
6. Modern Transformer Variants & BERT	6 Qs
7. LLM Fundamentals	11 Qs
8. Hugging Face Transformers	2 Qs
9. Prompt Engineering & Generation Parameters	9 Qs
PART C — BUILDING GenAI APPLICATIONS	
10. LangChain & LCEL	7 Qs
11. RAG & Vector Databases	17 Qs
12. Agents: Tools, LangGraph, CrewAI, MCP	14 Qs
13. Fine-tuning LLMs	6 Qs
14. Evaluation of LLMs & RAG	5 Qs
15. LLMOps, Deployment & Safety	8 Qs
PART D — INTERVIEW SCENARIOS	
16. System Design & Scenario Questions	7 Qs
17. Project & Behavioural Questions	4 Qs
PART E — ENGINEERING ESSENTIALS	
18. Python & ML Essentials for GenAI Engineers	8 Qs

How to use: answer each one out loud in under 2 minutes. Every question shows where it lives in the Assignments folder (e.g. Assignment 11 · Q3), and the full answer is in GenAI_Must_Know_Interview_Questions_Solutions.pdf. Company names show where the question was asked in real interviews.

PART A — FOUNDATIONS

1. NLP Basics & Text Preprocessing

1. What is NLP? Give 3 real-world applications you have seen in production. (Assignment 1 · Q1) ★
2. What are the main challenges in NLP? (Assignment 1 · Q3) ★
3. What is tokenization? Compare word-level, character-level and subword tokenization. (Assignment 1 · Q8) ★
4. What is stemming and how is it different from lemmatization? When would you prefer each? (Assignment 1 · Q9) ★ · Asked: NLP interview rounds
5. What are subword tokenization methods? Explain BPE, WordPiece and SentencePiece briefly. (Assignment 1 · Q15) ★
6. Which tokenizers are used by GPT-family models and by BERT? (Assignment 1 · Q17) ★
7. In real-time data users type "Lunch", "lunch" and "LUNCH". Why do we lowercase text, and when should you NOT lowercase? (Assignment 1 · Q22) ★ · Asked: NLP interview rounds
8. Users will type stopwords at inference time. If you remove them during training, what happens? When should you remove stopwords and when not? Give practical applications. (Assignment 1 · Q24) ★ · Asked: NLP interview rounds

2. Text Representation & Word Embeddings

1. What is the Bag-of-Words model? What are its limitations? (Assignment 2 · Q2) ★
2. What is TF-IDF? How do you calculate TF, IDF and TF-IDF? Work out a small example by hand. (Assignment 2 · Q4) ★ · Asked: NLP interview rounds
3. What are word embeddings? How are they better than one-hot / BoW vectors? (Assignment 2 · Q5) ★
4. Explain Word2Vec and its two architectures (CBOW and Skip-gram). What is the difference? (Assignment 2 · Q7) ★ · Asked: NLP interview rounds
5. How do you handle out-of-vocabulary (OOV) words? (Assignment 2 · Q11) ★
6. Explain cosine similarity and why it is important in NLP. (Assignment 2 · Q12) ★
7. Euclidean distance vs cosine similarity — when would you use which? (Assignment 2 · Q14) ★
8. Why can averaging Word2Vec vectors to make a sentence vector lose information? What is better? (Assignment 2 · Q20) ★ · Asked: NLP interview rounds
9. What are Sentence Transformers (SBERT)? Why are they better for semantic search than averaged word vectors or BERT's raw [CLS] vector? (Assignment 2 · Q21) ★ · Asked: NLP interview rounds
10. Explain embeddings to a non-technical person without using the word "vector". (Assignment 2 · Q23) ★ · Asked: Deloitte

3. Neural Network Fundamentals

1. How does a neural network work? Explain with a single neuron first. (Assignment 3 · Q1) ★
2. Explain the learning rate. What happens if it is too high or too low? (Assignment 3 · Q5) ★
3. Explain forward propagation and backpropagation. (Assignment 3 · Q6) ★
4. What are the different types of optimizers (SGD, Momentum, RMSProp, Adam, AdamW)? (Assignment 3 · Q8) ★
5. What is the use of an activation function? What happens without one? (Assignment 3 · Q13) ★
6. Which loss function should be used for regression, binary classification, multi-class and multi-label problems? (Assignment 3 · Q16) ★
7. What are vanishing and exploding gradients? Why do they occur and how do you fix them? (Assignment 3 · Q17) ★
8. If the previous layer has 8 neurons and the current Dense layer has 512 neurons, how many parameters does the current layer have? (Show the calculation.) (Assignment 3 · Q22) ★

4. Sequence Models: RNN, LSTM, GRU

1. What is the difference between an RNN and an ANN? (Assignment 4 · Q1) ★
2. Why do RNNs suffer from vanishing gradients and how can it be solved? (Assignment 4 · Q6) ★ · Asked: NLP interview rounds
3. Explain the LSTM cell: cell state, hidden state and the three gates (forget, input, output). (Assignment 4 · Q7) ★ · Asked: NLP interview rounds
4. How does LSTM address vanishing gradients and long-term dependencies? (Assignment 4 · Q9) ★
5. What is a GRU? Explain its gates (update, reset). How does it differ from LSTM? (Assignment 4 · Q11) ★
6. Why did Transformers replace RNNs/LSTMs for most NLP tasks? (Assignment 4 · Q14) ★

PART B — TRANSFORMERS & LLMs

5. Attention & Transformer Architecture

1. What is attention and why was it needed (the bottleneck problem)? (Assignment 5 · Q3) ★
2. What is self-attention? Explain Query, Key and Value with an intuitive example. (Assignment 5 · Q5) ★ · Asked: Google, NLP rounds
3. What is Scaled Dot-Product Attention? Write the formula. Why do we divide by $\sqrt{d_k}$? (Assignment 5 · Q6) ★
4. What is multi-head attention and why is it better than a single head? (Assignment 5 · Q7) ★
5. How do Transformers account for word order? Explain positional encoding using sine and cosine. (Assignment 5 · Q9) ★ · Asked: NLP interview rounds

6. What is masked (causal) attention and why is it used in the decoder? (Assignment 5 · Q14) ★
7. Explain the complete Transformer architecture end-to-end (draw it). (Assignment 5 · Q17) ★ · Asked: TCS, Google
8. What is the time and memory complexity of self-attention with respect to sequence length? (Assignment 5 · Q18) ★
9. How do LLMs understand words with multiple meanings, e.g. "bank" of a river vs a financial "bank"? (Assignment 5 · Q21) ★ · Asked: EPAM

6. Modern Transformer Variants & BERT

1. Give examples of encoder-only, decoder-only and encoder–decoder models. Which tasks suit each? (Assignment 6 · Q1) ★ · Asked: TCS
2. Explain Multi-Head Attention (MHA), Multi-Query Attention (MQA) and Grouped-Query Attention (GQA). (Assignment 6 · Q3) ★
3. What is the KV cache and why is it needed? How does it affect memory at long context lengths? (Assignment 6 · Q5) ★
4. What are the types of positional embeddings (absolute, learned, relative, RoPE, ALiBi)? Absolute vs relative? (Assignment 6 · Q8) ★
5. What is BERT? How is it bidirectional? (Assignment 6 · Q13) ★ · Asked: NLP interview rounds
6. Which tasks are used for BERT pre-training (MLM and NSP)? Why did RoBERTa drop NSP? (Assignment 6 · Q14) ★

7. LLM Fundamentals

1. What is a Large Language Model? How is it trained (pre-training → instruction tuning → alignment)? (Assignment 7 · Q1) ★ · Asked: Google
2. What is the difference between a base model and an instruct/chat model? (Assignment 7 · Q2) ★
3. What is RLHF? Explain the reward model and PPO. What is DPO and why is it simpler? (Assignment 7 · Q3) ★ · Asked: Google
4. What is a context window? What problems appear with very long contexts ("lost in the middle", cost, latency)? (Assignment 7 · Q5) ★
5. Explain decoding strategies: greedy, beam search, top-k, top-p (nucleus) and temperature. (Assignment 7 · Q6) ★ · Asked: TCS, EPAM, Google
6. What is quantization (INT8, 4-bit, GPTQ, AWQ, GGUF)? What do you trade off? (Assignment 7 · Q7) ★
7. What is hallucination in an LLM? What causes it and how do you reduce it? (Assignment 7 · Q12) ★ · Asked: TCS, Infosys, EPAM, Google
8. What is multimodality? Name a few multimodal models and explain how images are fed to them. (Assignment 7 · Q15) ★ · Asked: Google
9. What is a token? Roughly how many tokens is a word or a page, and why does token count matter? (Assignment 7 · Q19) ★ · Asked: Google

10. Can hallucinations be completely eliminated from a GenAI application? (Assignment 7 · Q20) ★ · Asked: Deloitte, Google

11. How would you choose an LLM for a production application, and how do you compare two LLMs? (Assignment 7 · Q22) ★ · Asked: Google

8. Hugging Face Transformers

1. Explain the Trainer and TrainingArguments classes. (Assignment 8 · Q11) ★

2. What are autoregressive models? (Assignment 8 · Q12) ★

9. Prompt Engineering & Generation Parameters

1. What are system, user and assistant messages? (SystemMessage, HumanMessage, AIMessage) (Assignment 9 · Q2) ★ · Asked: TCS, Infosys

2. Explain zero-shot, one-shot and few-shot (N-shot) prompting. (Assignment 9 · Q3) ★

3. What is Chain-of-Thought (CoT) prompting? When does it help? (Assignment 9 · Q4) ★

4. How do temperature, top_p and top_k affect generation? Which settings would you use for a factual Q&A bot vs creative writing? (Assignment 9 · Q9) ★ · Asked: TCS, EPAM, Google

5. How do you get reliable structured output (JSON) from an LLM? (Structured outputs, function calling, output parsers.) (Assignment 9 · Q12) ★ · Asked: Wipro, Google

6. What is prompt injection? How do you defend against it (including indirect injection via retrieved documents)? (Assignment 9 · Q13) ★ · Asked: Wipro, Google

7. What happens when temperature is set to 0? What if it is increased to 1.2? Is temperature 0 fully deterministic? (Assignment 9 · Q15) ★ · Asked: Deloitte, Google

8. What is context engineering? How is it different from prompt engineering? (Assignment 9 · Q16) ★ · Asked: Google

9. Should you always give more context to an LLM? What is context-window overflow and how do you handle it? (Assignment 9 · Q17) ★ · Asked: Google

PART C — BUILDING GenAI APPLICATIONS

10. LangChain & LCEL

1. What is LangChain and how does it help build LLM applications? What are its main components? (Assignment 10 · Q1) ★

2. What is LCEL (LangChain Expression Language)? Build a chain: prompt | model | output_parser. (Why did LCEL replace the legacy LLMChain?) (Assignment 10 · Q4) ★

3. What is the difference between invoke, stream and batch? (Assignment 10 · Q5) ★

4. How do you split documents? Explain RecursiveCharacterTextSplitter, chunk_size and chunk_overlap. Split a document into 1000-character chunks. (Assignment 10 · Q11) ★

5. How do you summarize a document larger than the context window? Explain stuff, map_reduce and refine. (Assignment 10 · Q12) ★

6. How do you make an LLM remember previous conversations? (Assignment 10 · Q13) ★

7. How do you handle long conversation memory without exceeding the context window or inflating token cost? (Assignment 10 · Q18) ★ · Asked: Wipro

11. RAG & Vector Databases

1. What is RAG? Explain its architecture (indexing pipeline + retrieval + generation) with a diagram. (Assignment 11 · Q1) ★ · Asked: TCS, PwC, Infosys, Google
2. RAG vs fine-tuning vs long-context prompting — when would you use each? (Assignment 11 · Q2) ★ · Asked: TCS, Infosys, Google
3. What are embeddings? Name OpenAI embedding models and open-source embedding models. Can you reduce OpenAI embedding dimensions? (Assignment 11 · Q3) ★ · Asked: Google
4. What are chunking strategies (fixed, recursive, semantic, document-structure based)? How do you choose chunk size? (Assignment 11 · Q5) ★ · Asked: TCS, Deloitte, Google
5. How are indexes created in vector databases? Explain flat, IVF, PQ and HNSW. (Assignment 11 · Q6) ★
6. How does HNSW (Hierarchical Navigable Small Worlds) work? (Assignment 11 · Q7) ★
7. What is MMR (Maximal Marginal Relevance) and why use it? (Assignment 11 · Q11) ★ · Asked: Wipro
8. What is hybrid search (BM25 + vector)? Why does it help? (Assignment 11 · Q12) ★ · Asked: Google
9. What is re-ranking? Explain cross-encoders and Cohere Rerank. (Assignment 11 · Q13) ★ · Asked: Wipro, Google
10. What is Agentic RAG? What is GraphRAG? (Assignment 11 · Q19) ★ · Asked: Google
11. Your RAG system gives wrong answers. How do you debug it step by step (retrieval vs generation)? (Assignment 11 · Q20) ★ · Asked: Deloitte, Wipro, Google
12. If frontier LLMs have huge context windows, why not send the entire PDF (or all documents) to the model instead of building RAG? (Assignment 11 · Q22) ★ · Asked: Deloitte, Google
13. What is the difference between tokenization and chunking? (Assignment 11 · Q23) ★ · Asked: PwC
14. How do you handle restricted vs unrestricted documents in RAG, so unauthorized documents never enter the LLM context? (Assignment 11 · Q25) ★ · Asked: TCS, EY, Google
15. How do you implement citations in RAG? And what if the answer is correct but the cited source is wrong? (Assignment 11 · Q26) ★ · Asked: Google
16. What happens if no relevant document is retrieved? How do you reduce hallucination when retrieval is weak? (Assignment 11 · Q27) ★ · Asked: Google
17. 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. (Assignment 11 · Q30) ★ · Asked: Deloitte, TCS

12. Agents: Tools, LangGraph, CrewAI, MCP

1. What is an AI agent? How is it different from a chain? (Assignment 12 · Q1) ★ · Asked: EPAM, PwC, Google
2. What are tools and how do they improve an LLM's capability? How does tool/function calling work? (Assignment 12 · Q2) ★ · Asked: TCS, Google
3. How does a ReAct agent work (Thought → Action → Observation loop)? (Assignment 12 · Q3) ★ · Asked: Google
4. What is LangGraph? Why use it instead of a simple agent loop? (Assignment 12 · Q8) ★ · Asked: Google
5. What is CrewAI? How do you create agents, tasks and run a crew? (Assignment 12 · Q12) ★
6. What is MCP (Model Context Protocol) and what problem does it solve? (Assignment 12 · Q14) ★ · Asked: TCS, EPAM, Infosys, Google
7. What is Agentic AI, and how is it different from standard Generative AI or a RAG chatbot? (Assignment 12 · Q17) ★ · Asked: TCS, Infosys, PwC
8. Who decides which tool to call? What if the model selects the wrong tool? Why shouldn't an agent have access to every tool? (Assignment 12 · Q18) ★ · Asked: Infosys, Google
9. Your agent has 10 tools. The user asks a simple question, but the agent calls 5 different tools. How do you fix this? (Assignment 12 · Q19) ★ · Asked: Infosys
10. Would you let an LLM perform a sensitive action automatically, like deleting a customer record or making a financial transaction? Should the LLM decide permissions? (Assignment 12 · Q20) ★ · Asked: Infosys, Google
11. What is the difference between normal function calling and MCP? In MCP, what is the difference between tools and resources? (Assignment 12 · Q21) ★ · Asked: Infosys, EPAM
12. How would you run three independent agents in parallel and make sure one agent's failure doesn't break the others? How do you add retries with backoff? (Assignment 12 · Q22) ★ · Asked: EPAM
13. LangChain vs LangGraph – when would you use each? (Assignment 12 · Q23) ★ · Asked: Google
14. In a Text-to-SQL agent, how do you prevent the LLM from executing destructive or hallucinated SQL queries? (Assignment 12 · Q26) ★ · Asked: Deloitte

13. Fine-tuning LLMs

1. When should you fine-tune an LLM and when should you not? (Assignment 13 · Q1) ★ · Asked: Infosys, Google
2. What is Parameter-Efficient Fine-Tuning (PEFT) and why is it needed? (Assignment 13 · Q2) ★ · Asked: TCS
3. What are the challenges of full fine-tuning and how does LoRA overcome them? (Assignment 13 · Q3) ★
4. Explain how LoRA works. How are the weight matrices updated? (W + BA) (Assignment 13 · Q4) ★

5. LoRA vs QLoRA — what is the difference and when would you use each? (Assignment 13 · Q6) ★
6. What are the end-to-end steps to fine-tune an LLM for a business use case? (Assignment 13 · Q14) ★ · Asked: PwC

14. Evaluation of LLMs & RAG

1. How do you evaluate an LLM application? (Offline vs online, reference-based vs reference-free.) (Assignment 14 · Q1) ★ · Asked: Infosys, Google
2. What is RAGAS? What metrics evaluate the retriever and the generator? (Assignment 14 · Q2) ★ · Asked: Google
3. Explain faithfulness, answer relevancy, context precision and context recall. (Assignment 14 · Q3) ★
4. How do you evaluate retrieval quality? Explain Precision@K, Recall@K, MRR (and nDCG). (Assignment 14 · Q9) ★ · Asked: TCS
5. How do you compare two prompt versions, and what are regression testing and shadow testing for LLM apps? (Assignment 14 · Q10) ★ · Asked: Google

15. LLMOps, Deployment & Safety

1. How do you deploy an LLM application to production? Describe the architecture. (Assignment 15 · Q1) ★ · Asked: Google
2. How do you reduce latency and cost (caching, smaller models, streaming, batching, prompt compression, routing)? (Assignment 15 · Q2) ★ · Asked: TCS, Google
3. What are guardrails? How do you handle PII, toxic content and jailbreaks? (Assignment 15 · Q6) ★ · Asked: Google
4. What is the difference between a POC and a production-grade GenAI architecture? (Assignment 15 · Q10) ★ · Asked: PwC
5. How would you host an LLM application to keep latency low during high traffic spikes? (Assignment 15 · Q11) ★ · Asked: TCS, Google
6. How do you handle flaky LLM APIs – timeouts, rate limits (429) and server errors? (Assignment 15 · Q12) ★ · Asked: Wipro, Google
7. What is prompt caching, and how is it different from semantic caching? How do you invalidate a cache? (Assignment 15 · Q13) ★ · Asked: Google, Wipro
8. What is the OWASP Top 10 for LLM applications? Explain excessive agency and data exfiltration. (Assignment 15 · Q14) ★ · Asked: Google

PART D — INTERVIEW SCENARIOS

16. System Design & Scenario Questions

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. (Assignment 16 · Q1) ★ · Asked: Deloitte, EY, Google

2. Your RAG chatbot answers correctly in testing but users complain in production. How do you investigate? (Assignment 16 · Q2) ★ · Asked: Google
3. 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? (Assignment 16 · Q11) ★ · Asked: EY
4. The knowledge base grew from 10,000 to 5 million documents. Retrieval latency crashed and hosting cost spiked. What do you change? (Assignment 16 · Q12) ★ · Asked: EY
5. A non-technical client insists the chatbot must be "100% accurate and never make a mistake". How do you handle this without jargon? (Assignment 16 · Q13) ★ · Asked: EY, Deloitte, Google
6. Production responses suddenly became worse right after a deployment. What do you investigate? (Assignment 16 · Q14) ★ · Asked: Google
7. Design a system that processes and summarizes 10 million documents. (Assignment 16 · Q16) ★ · Asked: Google

17. Project & Behavioural Questions

1. Walk me through a GenAI project you built end-to-end. What was your role and what was the business impact? (Assignment 17 · Q1) ★ · Asked: TCS, Deloitte, Infosys, Google
2. What was the hardest problem you faced in that project and how did you solve it? (Assignment 17 · Q2) ★ · Asked: Google
3. Tell me about yourself (for a GenAI Engineer role). (Assignment 17 · Q7) ★ · Asked: Infosys, Google
4. Tell me about a failure or production issue in your GenAI project and what you learned. (Assignment 17 · Q8) ★ · Asked: Google

PART E — ENGINEERING ESSENTIALS

18. Python & ML Essentials for GenAI Engineers

1. What is the difference between multithreading and multiprocessing in Python? What is the GIL, and how does it affect GenAI backends? (Assignment 18 · Q1) ★ · Asked: TCS, Infosys
2. Concurrency vs parallelism: why use async programming in an AI application? If you must call three independent APIs, would you call them one by one? (Assignment 18 · Q2) ★ · Asked: Infosys
3. What are generators and decorators in Python? How do generators relate to streaming LLM responses? (Assignment 18 · Q3) ★ · Asked: TCS, Google
4. Why is FastAPI preferred for GenAI backends, and how does Pydantic fit in? (Assignment 18 · Q7) ★ · Asked: EPAM, Google
5. Precision vs recall vs F1 score – when is recall more important than precision? (Assignment 18 · Q11) ★ · Asked: Google
6. What is overfitting, and what is data leakage? How do you prevent each? (Assignment 18 · Q12) ★ · Asked: Google

7. How do you handle class imbalance (e.g. 1% fraud vs 99% normal transactions)? (Assignment 18 · Q13) ★ · Asked: NLP interview rounds, Google

8. What is data drift vs model (concept) drift vs training–serving skew? Batch vs online inference? (Assignment 18 · Q16) ★ · Asked: Google