Generative AI Interview Q&A Guide - From Beginner to Expert

Prepared by Analytics Circle

This document contains the first 30 detailed Generative AI interview questions and answers, covering beginner and intermediate levels. Each question includes detailed explanations to help you understand concepts clearly for interviews.

1. What is Generative AI?

Answer: Generative AI is a branch of artificial intelligence that creates new, original data such as text, images, music, or code that mimic human-created content. It leverages models like GANs, VAEs, and Transformers to learn from large datasets and produce realistic, contextually relevant outputs.

2. How is Generative AI different from Traditional AI?

Answer: Traditional AI performs classification or prediction tasks, while Generative AI focuses on creativity by generating new content like essays, designs, or art.

3. What are the main applications of Generative AI?

Answer: Applications include text generation (ChatGPT), image synthesis (DALL-E), coding assistance (Copilot), summarization, and chatbot development.

4. What are Large Language Models (LLMs)?

Answer: LLMs are deep learning models trained on massive datasets to understand and generate natural language. Examples include GPT-4, Claude, Gemini, and LLaMA.

5. What is the Transformer architecture?

Answer: Transformers use self-attention to capture contextual relationships in sequences, replacing recurrent networks (RNNs) for efficiency and scalability.

6. What is the Attention mechanism?

Answer: Attention assigns importance weights to words, enabling the model to focus on relevant parts of the input while generating or interpreting text.

7. Difference between BERT and GPT?

Answer: BERT is encoder-only and bidirectional for understanding tasks, while GPT is decoder-only and unidirectional for text generation tasks.

8. What are tokens and tokenization?

Answer: Tokenization breaks text into smaller units (tokens) for processing. Example: 'ChatGPT rocks!' becomes ['Chat', 'G', 'PT', 'rocks', '!'].

9. What is temperature in LLMs?

Answer: Temperature controls randomness in text generation. Low values make outputs deterministic; higher values make them creative and diverse.

10. What is fine-tuning?

Answer: Fine-tuning retrains a pre-trained model on smaller, domain-specific datasets for customized applications like legal or healthcare assistants.

11. What are GANs?

Answer: Generative Adversarial Networks consist of a generator and a discriminator that compete, resulting in highly realistic synthetic data such as deepfake images.

12. What is a Variational Autoencoder (VAE)?

Answer: A VAE encodes input into a latent space and reconstructs it, learning distributions that allow generating new, similar samples.

13. GAN vs VAE?

Answer: GANs produce sharper outputs via adversarial training but are unstable. VAEs are stable but generate blurrier outputs.

14. What is prompt engineering?

Answer: Prompt engineering is crafting effective inputs that guide models to produce relevant, accurate responses by including context, examples, or roles.

15. What is RAG?

Answer: Retrieval-Augmented Generation retrieves relevant data from a knowledge base and combines it with generation for fact-based answers.

16. Common challenges in GenAI?

Answer: Challenges include hallucination, bias, privacy concerns, and high computational requirements.

17. What is RLHF?

Answer: Reinforcement Learning from Human Feedback aligns model responses with human intent using reward modeling and fine-tuning.

18. Evaluation metrics for LLMs?

Answer: Metrics include BLEU, ROUGE, FID, perplexity, and human evaluation for relevance and accuracy.

19. What are embeddings?

Answer: Embeddings represent text or data as numerical vectors capturing semantic relationships for similarity searches.

20. What is a vector database?

Answer: Databases like Pinecone or FAISS store embeddings and support fast semantic retrieval in RAG

systems.

21. What is prompt chaining?

Answer: A technique where multiple prompts are linked so that the output of one becomes input for another to achieve complex reasoning.

22. Few-shot vs zero-shot learning?

Answer: Few-shot provides task examples within prompts; zero-shot expects the model to infer the task without examples.

23. What is LoRA?

Answer: Low-Rank Adaptation fine-tunes only small trainable matrices, reducing compute cost while retaining model performance.

24. What is quantization?

Answer: Quantization reduces numerical precision (e.g., FP32 to INT8), lowering memory use and speeding up inference.

25. What is grounding?

Answer: Grounding connects generated text to factual data sources to ensure responses are accurate and trustworthy.

26. What is chain-of-thought prompting?

Answer: Encourages step-by-step reasoning by prompting the model to explain intermediate steps in logical tasks.

27. What are multimodal models?

Answer: Models like GPT-4V or Gemini process text, images, and audio together for comprehensive understanding.

28. What are diffusion models?

Answer: Diffusion models start with random noise and iteratively refine it into realistic images, used in Stable Diffusion.

29. What is prompt tuning?

Answer: Prompt tuning learns optimized prompts (soft tokens) to guide model behavior without changing model weights.

30. Fine-tuning vs prompt tuning?

Answer: Fine-tuning retrains model weights; prompt tuning adjusts prompts, making it faster and more cost-efficient.