Generative AI Interview Q&A Guide - Intermediate to Expert Level

Prepared by Analytics Circle

This document contains the next 35 detailed Generative AI interview questions and answers covering intermediate to expert-level topics such as fine-tuning, LoRA, RAG, RLHF, quantization, and enterprise deployment strategies.

31. What is PEFT (Parameter Efficient Fine-Tuning)?

Answer: PEFT fine-tunes only a small portion of model parameters such as adapters or low-rank matrices, instead of all weights. This reduces GPU memory requirements and training costs while maintaining performance.

32. What is RAG (Retrieval-Augmented Generation)?

Answer: RAG retrieves relevant information from a database and combines it with generation to create factually accurate answers. It is widely used for chatbots and domain-specific AI assistants.

33. What is RLHF (Reinforcement Learning from Human Feedback)?

Answer: RLHF aligns models with human preferences through three stages: supervised fine-tuning, reward modeling, and policy optimization (using PPO). It ensures responses are safe and contextually relevant.

34. What are embeddings used for in LLMs?

Answer: Embeddings represent text as numerical vectors, capturing semantic meaning. They enable search, clustering, and similarity-based retrieval.

35. Difference between pre-training and fine-tuning?

Answer: Pre-training teaches models general patterns using massive data, while fine-tuning adapts them for specific tasks like sentiment analysis or summarization.

36. What are diffusion models?

Answer: Diffusion models generate images by iteratively denoising random noise, learning to reverse the noise process during training. Used in Stable Diffusion and DALL-E 3.

37. What is LoRA (Low-Rank Adaptation)?

Answer: LoRA inserts small trainable layers into frozen pre-trained weights, reducing fine-tuning cost and enabling modular domain updates.

38. What is quantization?

Answer: Quantization reduces numerical precision (e.g., FP32 INT8), cutting model size and inference latency with minimal accuracy loss.

39. What is pruning?

Answer: Pruning removes redundant neurons or connections from neural networks to improve efficiency without hurting accuracy.

40. What are hallucinations in LLMs?

Answer: Hallucinations occur when models generate false or unverifiable content. They can be reduced using RAG, factual grounding, and temperature control.

41. What are multimodal LLMs?

Answer: Models like GPT-4V and Gemini can process multiple input types (text, image, audio, video) for richer interaction and reasoning.

42. What are attention heads?

Answer: Attention heads in Transformers learn different aspects of relationships between tokens, enabling multi-context understanding.

43. What is context window in LLMs?

Answer: It defines the maximum number of tokens the model can process at once. GPT-4 Turbo supports up to 128K tokens.

44. Difference between prompt tuning and instruction tuning?

Answer: Prompt tuning adjusts input prompts using soft tokens, while instruction tuning fine-tunes models on labeled instruction datasets.

45. What is chain-of-thought prompting?

Answer: A prompting technique that guides the model to reason step-by-step, improving logical and mathematical accuracy.

46. What are vector databases?

Answer: Databases like FAISS and Pinecone store embeddings for fast, similarity-based retrieval in AI pipelines.

47. Explain a RAG pipeline architecture.

Answer: A RAG pipeline has a retriever that fetches relevant data and a generator that uses it for grounded text generation. Often implemented with LangChain.

48. What are the main challenges in deploying LLMs at scale?

Answer: Challenges include high latency, scalability, cost, data privacy, and model drift. Solutions involve quantization, caching, and distributed inference.

49. How do you evaluate fine-tuned LLMs?

Answer: Evaluation uses BLEU, ROUGE, F1, perplexity, and human scoring for factuality and fluency.

50. Explain model quantization and pruning in optimization.

Answer: Quantization reduces precision; pruning removes unnecessary neurons. Both improve model efficiency.

51. What is grounding in LLMs?

Answer: Grounding ties generated outputs to factual or real-world sources, ensuring trust and factual correctness.

52. What are embeddings used for enterprise GenAI?

Answer: They enable enterprise chatbots and AI tools to perform semantic search on private data repositories efficiently.

53. What is RLHF and why is it important?

Answer: It ensures AI aligns with human values, improving safety, tone, and preference adherence during conversation.

54. What is the difference between fine-tuning and instruction tuning?

Answer: Fine-tuning adapts to domain tasks; instruction tuning trains on diverse instructions to improve general usability.

55. What are the key challenges in GenAI?

Answer: Bias, hallucination, copyright issues, interpretability, and high compute cost.

56. How does fine-tuning differ by architecture?

Answer: Encoder-only models are fine-tuned for understanding tasks; decoder-only for generation; encoder-decoder for translation or summarization.

57. How does RAG improve accuracy?

Answer: RAG retrieves relevant documents before generation, grounding model output and reducing hallucinations.

58. What is LoRA and how does it reduce fine-tuning cost?

Answer: It trains small low-rank matrices while keeping original weights frozen, cutting training cost by 90%+.

59. How to evaluate fine-tuned models?

Answer: Use automatic (BLEU, ROUGE) and human evaluation for relevance, coherence, and truthfulness.

60. What are challenges in LLM deployment?

Answer: Latency, scalability, cost, data compliance, and version drift.

61. Explain quantization vs pruning.

Answer: Quantization reduces numerical precision, pruning removes redundant parameters to shrink model size.

62. What is RLHF alignment?

Answer: RLHF aligns LLMs with human values using human preference data and reinforcement learning

optimization.

63. Fine-tuning vs instruction tuning?

Answer: Fine-tuning = domain specialization; instruction tuning = better generalization for user intent.

64. How to ensure data privacy in GenAI projects?

Answer: Use on-prem models, encrypt sensitive data, anonymize PII, and comply with GDPR or HIPAA regulations.

65. Explain an enterprise-grade RAG pipeline for AI assistants.

Answer: Involves document ingestion, embedding storage in vector DB, retriever, and generator LLM integration for factual responses with logging.