Top 100 Gen AI Engineer Interview Questions

1. Fundamentals of Generative AI

1. **What is Generative AI?**
   Generative AI creates new data (text, image, code, etc.) using models like LLMs or diffusion models trained on massive datasets.

2. **Difference between Predictive AI and Generative AI?**
   Predictive AI → forecasts outcomes.
   Generative AI → creates new content.

3. **What are LLMs?**
   Large Language Models trained on billions of tokens to generate human-like text.

4. **What is tokenization?**
   Converting text into smaller units (tokens) that the model understands.

5. **What is a transformer architecture?**
   A neural network using self-attention to process sequences in parallel.

6. **Difference between GPT, BERT, and T5?**
   - BERT → encoder, bidirectional, for understanding tasks.
   - GPT → decoder, autoregressive, for generation.
   - T5 → encoder-decoder, versatile for many tasks.

7. **What is "context window"?**
   Maximum number of tokens the model can process at once.

8. **What is fine-tuning vs. pre-training?**
   Pre-training = on massive general data.
   Fine-tuning = adapting to domain-specific tasks.

9. **Explain embeddings in LLMs.**
   Vector representations of text used for similarity, search, or retrieval.

10. **What is RLHF (Reinforcement Learning from Human Feedback)?**
    Aligning AI responses with human preferences via feedback-driven reinforcement.

2. Prompt Engineering

11. **What is prompt engineering?**
    Crafting inputs to guide LLMs for accurate outputs.

12. **Difference between zero-shot, one-shot, and few-shot prompting?**

- Zero-shot → no examples.
- One-shot → one example.
- Few-shot → multiple examples.

13. **What is chain-of-thought prompting?**
    Asking model to show reasoning steps before giving final answer.

14. **What is prompt injection attack?**
    Malicious inputs altering AI's intended behavior.

15. **How do you avoid hallucinations in LLMs?**
    Use retrieval augmentation, fact-checking, and controlled prompts.

16. **What is system vs. user vs. assistant prompt?**

- System → defines role.
- User → request.
- Assistant → response.

17. **What is self-consistency in prompting?**
    Asking model to generate multiple answers and picking the most consistent one.

18. **What is prompt chaining?**
    Breaking complex tasks into multiple prompt steps.

19. **What is role prompting?**
    Assigning a role to model (e.g., "You are a math teacher").

20. **What is few-shot CoT prompting?**
    Combining few examples with reasoning steps for complex problems.

3. Workflows & Automation

21. **What is a GenAI workflow?**
    A pipeline where LLMs interact with data sources, APIs, and tools for automation.

22. **What tools are used for orchestration?**
    LangChain, LlamaIndex, Haystack, Prefect, Airflow.

23. **What is Retrieval-Augmented Generation (RAG)?**
    LLM retrieves relevant docs (via vector DB) + generates response.

24. **How do you automate document processing with LLMs?**
    Ingest docs → vectorize → query via RAG → summarize or answer.

25. **What are vector databases?**
    Databases (Pinecone, Weaviate, Milvus, FAISS) that store embeddings for similarity search.

26. **What's the role of APIs in GenAI automation?**
    They allow LLMs to integrate with external services (Slack, CRMs, databases).

27. **What is function calling in LLMs?**
    When LLMs call external functions/APIs to complete tasks.

28. **What's the difference between pipelines vs. agents?**

- Pipeline → fixed workflow.

- Agent → autonomous, flexible decision-making.

29. **How do you automate workflows in enterprises?**
    Combine LLMs with RPA tools (UiPath, Automation Anywhere, Make, Zapier).

30. **What is an AI-powered knowledge assistant workflow?**
    Input query → embedding search → context retrieval → LLM response.

4. Agentic AI

31. **What is Agentic AI?**
    AI agents that autonomously plan, reason, and execute tasks using tools.

32. **Difference between LLMs and AI agents?**
    LLMs → text generation only.
    Agents → LLM + memory + tool use + decision-making.

33. **What is AutoGPT?**
    An autonomous agent framework using GPT to plan multi-step tasks.

34. **What is ReAct prompting?**
    Reason + Act framework for LLMs to use tools step by step.

35. **What is multi-agent system?**
    Multiple AI agents collaborating to solve tasks.

36. **How do agents use memory?**
    By storing conversation states or embeddings for context.

37. **What is tool use in agents?**
    Calling APIs, databases, or calculators to complete tasks.

38. **What's the difference between reflexive vs. deliberative agents?**
    Reflexive = immediate response.
    Deliberative = reasoning before action.

39. **What are some popular agent frameworks?**
    LangChain Agents, CrewAI, OpenAI Functions, AutoGPT, BabyAGI.

40. **What's the biggest challenge in agentic AI?**
    Hallucination, reliability, and safe tool execution.

5. Deployment & MLOps

41. **How do you deploy LLM apps?**
    Via REST APIs, Docker, Kubernetes, or cloud services (AWS Sagemaker, GCP Vertex AI).

42. **What is MLOps for LLMs?**
    Applying DevOps to ML → monitoring, versioning, retraining.

43. **What is model checkpointing?**
    Saving intermediate training states to resume later.

44. **What is A/B testing in GenAI apps?**
    Comparing outputs of two models/prompts to pick best one.

45. **How do you monitor GenAI models?**
    Track latency, token usage, accuracy, user feedback.

46. **What is model drift?**
    When model performance declines as data distribution changes.

47. **How do you scale LLM APIs?**
    Use load balancing, caching, batching requests.

48. **What is a vector cache?**
    Storing embeddings for repeated queries to reduce cost.

49. **What's the role of GPUs in LLMs?**
    Speed up training/inference with parallel computation.

50. **How do you handle cost optimization in LLM deployments?**
    Use smaller models, quantization, batching, and caching.

6. Data & Training

51. **What is synthetic data in GenAI?**
    AI-generated training data to augment real datasets.

52. **Difference between fine-tuning and LoRA?**
    LoRA = Low-Rank Adaptation, lightweight fine-tuning for efficiency.

53. **What's PEFT (Parameter-Efficient Fine-Tuning)?**
    Fine-tuning only small parts of model parameters.

54. **What is prompt-tuning?**
    Training only soft prompt embeddings instead of whole model.

55. **What is quantization?**
    Reducing model precision (FP32 → INT8) for faster inference.

56. **What is knowledge distillation?**
    Training a smaller model (student) from a larger one (teacher).

57. **What is catastrophic forgetting?**
    Fine-tuned model forgets pre-trained knowledge.

58. **What is instruction tuning?**
    Fine-tuning models to follow human-like instructions.

59. **What is multimodal training?**
    Training models on text, images, audio, and video.

60. **What is dataset curation for LLMs?**
    Filtering, deduplicating, cleaning raw data before training.

7. Security, Ethics & Governance

61. **What are hallucinations?**
    Confident but incorrect outputs by LLMs.

62. **How do you reduce hallucinations?**
    RAG, fact-checking, fine-tuning.

63. **What is AI bias?**
    When models reflect unfair patterns from training data.

64. **How do you mitigate AI bias?**
    Diverse datasets, fairness metrics, bias detection tools.

65. **What is adversarial prompting?**
    Maliciously crafted prompts to bypass safeguards.

66. **What is data leakage in AI?**
    Sensitive data unintentionally included in training or output.

67. **What is red-teaming in AI?**
    Testing models for vulnerabilities and unsafe outputs.

68. **What is model interpretability?**
    Ability to explain how a model makes decisions.

69. **What is GDPR compliance in AI?**
    Ensuring AI respects data privacy laws.

70. **What's the role of watermarking in GenAI?**
    Embedding hidden signals to detect AI-generated content.

8. Tools & Ecosystem

71. **What is LangChain?**
    A framework to build LLM-powered apps and agents.

72. **What is LlamaIndex?**
    Tool for connecting LLMs with external data sources.

73. **What is Hugging Face Transformers?**
    Library for pretrained NLP models.

74. **What is OpenAI Function Calling?**
    LLMs structured to call external APIs/functions.

75. **What is Pinecone?**
    A vector database for similarity search.

76. **What is Weaviate?**
    Open-source vector DB with semantic search.

77. **What is CrewAI?**
    Framework for orchestrating multiple AI agents.

78. **What is Haystack?**
    RAG pipeline framework.

79. **What is Rasa?**
    Framework for conversational AI bots.

80. **What is Guardrails AI?**
    Tool to add safety, validation, and guardrails to LLM responses.

9. Applications & Use Cases

81. **What is GenAI in customer support?**
    AI chatbots, virtual assistants, auto-replies.

82. **GenAI in content creation?**
    Blogs, ads, videos, scripts, images.

83. **GenAI in software development?**
    Code completion, debugging, automated testing.

84. **GenAI in healthcare?**
    Drug discovery, clinical trial summarization, patient records.

85. **GenAI in finance?**
    Fraud detection, report automation, investment research.

86. **GenAI in legal?**
    Contract analysis, compliance checks, legal summarization.

87. **What is an AI copilot?**
    AI assistant integrated into apps to support human tasks.

88. **What is autonomous research with GenAI?**
    Agents reading papers, summarizing, and generating insights.

89. **What is personalized learning with GenAI?**
    AI tutors adapting to student's style.

90. **What are multi-modal AI applications?**
    AI that understands text, speech, and vision together.

10. Future & Advanced Topics

91. **What is OpenAI's GPT-4o / multimodal LLM?**
    A model handling text, vision, and audio.

92. **What are SLMs (Small Language Models)?**
    Lightweight LLMs optimized for edge devices.

93. **What is federated learning in AI?**
    Training models on decentralized data (privacy-first).

94. **What is continual learning?**
    Models learning new info without forgetting old.

95. **What is self-improving AI?**
    Agents that refine their skills autonomously.

96. **What is AutoML in GenAI?**
    Automated ML pipeline design for model training.

97. **What are AI agents with memory?**
    Agents storing knowledge for long-term context.

98. **What is reasoning-aware AI?**
    AI capable of logical reasoning, not just pattern-matching.

99. **What's the difference between AGI & GenAI?**
    GenAI = narrow domain creativity.
    AGI = general intelligence like humans.

100. **Where is GenAI heading in next 5 years?**
    More **agentic**, multimodal, domain-specialized, and embedded across industries.