Top GenAI Interview Questions (2025 Edition) – X, Google, Microsoft, WhatsApp, and More

Table of Contents

1. Introduction

This document provides over 25 pages of the latest and most challenging interview questions and exercises used in hiring processes for top GenAI roles at companies like X (formerly Twitter), Google DeepMind, Microsoft, OpenAI, Meta, WhatsApp, and Anthropic.

2. Core GenAI Knowledge Questions

Q1: What is the difference between autoregressive and autoencoding models in NLP?

Q2: Compare BERT, GPT-4, Claude, and Gemini models.

Q3: Explain the concept of tokenization in large language models. What challenges do different tokenizers present?

Q4: How does attention differ from self-attention in Transformer models?

Q5: What are common pretraining objectives for GenAI models?

3. Prompt Engineering Challenges

Q6: Design a prompt to extract structured data from an unstructured legal document.

Q7: Rewrite this prompt to make it more robust against hallucinations: "Summarize this text in one paragraph."

Q8: How would you chain prompts for a multi-step workflow like document summarization and sentiment analysis?

Q9: Create an adversarial prompt that breaks the summarization guardrails of a basic LLM.

---

4. LLM Architecture and Fine-Tuning

Q10: Describe the full architecture of a Transformer-based decoder-only model.

Q11: Explain LoRA and why it's useful for fine-tuning GenAI models.

Q12: How do techniques like PEFT and QLoRA reduce memory usage in training?

Q13: Describe the pipeline to fine-tune an LLM using instruction tuning and RLHF.

---

5. Multimodal Model Integration

Q14: What are the challenges of integrating image and text inputs in a single model?

Q15: Describe how CLIP models align vision and language. How are they used in retrieval systems?

Q16: Design a multimodal input pipeline to process OCR + image captioning for a social app.

---

6. Evaluation Metrics and Debugging

Q17: How do you evaluate factual accuracy in generative models?

Q18: What's the difference between BLEU, ROUGE, and BERTScore?

Q19: Write a debugging plan for an LLM that's giving inconsistent outputs for the same input.

Q20: How can you detect and reduce hallucination in open-ended answers?

---

7. Ethics, Bias, and Safety in GenAI

Q21: How would you audit a model for gender or racial bias?

Q22: What is prompt injection? How can it be mitigated?

Q23: How do AI alignment and safety differ from traditional AI fairness principles?

Q24: Write an ethical risk analysis for a GenAI-powered hiring assistant.

---

8. System Design Questions

Q25: Design a scalable GenAI-powered chat system for WhatsApp.

Q26: What caching strategies would you use to reduce latency in LLM API calls?

Q27: Architect a secure prompt-routing backend for a multi-model GenAI assistant.

---

9. Real-World Case Study Questions

Q28: Meta wants to summarize Instagram DMs using LLMs. What approach would you take?

Q29: Google wants to build an LLM-based classroom assistant. List the pipeline stages from input to feedback.

Q30: Twitter (X) wants to auto-generate trending topics from tweets. Build the system pipeline.

---

10. Coding Exercises – Python & GenAI APIs

Q31: Use Hugging Face `transformers` to load and query a BERT model.

```
from transformers import pipeline
qa = pipeline("question-answering", model="distilbert-base-cased-distilled-squad")
qa({"question": "What's the capital of France?", "context": "Paris is the capital of France."})
```

Q32: Write a Python script that uses OpenAI API to summarize an article.

Q33: Write a tokenizer that splits text into overlapping n-grams.

Q34: Build a Flask API that takes a question and returns a GPT-4 answer.

---

11. API Design for GenAI Products

Q35: Design a RESTful API for prompt generation and storage.

Q36: How would you version and A/B test GenAI APIs in production?

Q37: Discuss the tradeoffs between serverless and containerized GenAI model inference.

---

12. Deployment and Scaling

Q38: What strategies would you use to deploy LLMs on edge devices?

Q39: How do you cache LLM embeddings at scale?

Q40: Compare vector DBs like Pinecone, Weaviate, and FAISS.

---

13. Research & Product Strategy

Q41: How would you pitch a GenAI assistant for enterprise documentation workflows?

Q42: How do you measure the success of a GenAI feature in a social media app?

Q43: Propose a roadmap for a multi-year GenAI research product.

---

14. Open-Ended Thought Exercises

Q44: Where do you see GenAI heading in the next five years?

Q45: What unsolved problems excite you most in the GenAI space?

Q46: Would you rather fine-tune a domain-specific model or use RAG? Why?

---

15. Additional Resources

- Hugging Face Course: https://huggingface.co/course
- OpenAI Cookbook: https://github.com/openai/openai-cookbook
- DeepLearning.AI's GenAI specialization
- Google's Model Garden and Gemini Playground
- Microsoft's Phi-3 and Azure AI Studio

End of Document

(Approx. 26 pages in formatted layout)