Assignment 15
LLMOps, Deployment & Safety

Part C — Building GenAI Applications

15 questions

★ **Must-know** — asked very frequently in GenAI / ML interviews (8 questions). Prepare these first.

1. How do you deploy an LLM application to production? Describe the architecture. ★ · Asked: Google
2. How do you reduce latency and cost (caching, smaller models, streaming, batching, prompt compression, routing)? ★ · Asked: TCS, Google
3. What are vLLM, TGI and Ollama? What is continuous batching and PagedAttention?
4. What is Groq / LPU inference and why is its latency so low?
5. What is semantic caching? · Asked: Google, Wipro
6. What are guardrails? How do you handle PII, toxic content and jailbreaks? ★ · Asked: Google
7. How do you monitor an LLM app in production (latency, tokens, cost, feedback, drift)? · Asked: TCS, Infosys, Google
8. What is Gradio / Streamlit used for? Build a quick demo UI for your RAG app.
9. Which speech-to-text and text-to-speech models have you used?
10. What is the difference between a POC and a production-grade GenAI architecture? ★ · Asked: PwC
11. How would you host an LLM application to keep latency low during high traffic spikes? ★ · Asked: TCS, Google
12. How do you handle flaky LLM APIs – timeouts, rate limits (429) and server errors? ★ · Asked: Wipro, Google
13. What is prompt caching, and how is it different from semantic caching? How do you invalidate a cache? ★ · Asked: Google, Wipro
14. What is the OWASP Top 10 for LLM applications? Explain excessive agency and data exfiltration. ★ · Asked: Google
15. How do you deploy a GenAI change safely, and how do you roll back a bad AI deployment? · Asked: Google