Assignment 7
LLM Fundamentals

Part B — Transformers & LLMs

22 questions

★ **Must-know** — asked very frequently in GenAI / ML interviews (11 questions). Prepare these first.

1. What is a Large Language Model? How is it trained (pre-training → instruction tuning → alignment)? ★ · Asked: Google
2. What is the difference between a base model and an instruct/chat model? ★
3. What is RLHF? Explain the reward model and PPO. What is DPO and why is it simpler? ★ · Asked: Google
4. What are scaling laws? What did the Chinchilla paper show?
5. What is a context window? What problems appear with very long contexts ("lost in the middle", cost, latency)? ★
6. Explain decoding strategies: greedy, beam search, top-k, top-p (nucleus) and temperature. ★ · Asked: TCS, EPAM, Google
7. What is quantization (INT8, 4-bit, GPTQ, AWQ, GGUF)? What do you trade off? ★
8. How much GPU memory is needed to load a 7B model in FP16? In 4-bit? (Show the calculation.)
9. What is knowledge distillation? How are small models trained from large ones?
10. What are emergent abilities and in-context learning?
11. What are reasoning models (models that "think" before answering)? When are they worth the extra latency and cost?
12. What is hallucination in an LLM? What causes it and how do you reduce it? ★ · Asked: TCS, Infosys, EPAM, Google
13. Open-source vs closed (API) LLMs — how do you choose for an enterprise use case (cost, privacy, quality, control)?
14. What is the difference between an LLM (text-completion) interface and a Chat model interface?
15. What is multimodality? Name a few multimodal models and explain how images are fed to them. ★ · Asked: Google
16. How are image generation (diffusion) models different from LLMs?
17. What is a GAN? Explain the roles of the generator and discriminator.
18. Generative vs discriminative models — explain with examples.

19. What is a token? Roughly how many tokens is a word or a page, and why does token count matter? ★ · Asked: Google

20. Can hallucinations be completely eliminated from a GenAI application? ★ · Asked: Deloitte, Google

21. What is a foundation model, and what does "alignment" mean for LLMs? · Asked: Google

22. How would you choose an LLM for a production application, and how do you compare two LLMs? ★ · Asked: Google