Assignment 9
Prompt Engineering & Generation Parameters

Part B — Transformers & LLMs

18 questions

★ **Must-know** — asked very frequently in GenAI / ML interviews (9 questions). Prepare these first.

1. What is a prompt? What is prompt engineering?
2. What are system, user and assistant messages? (SystemMessage, HumanMessage, AIMessage) ★ · Asked: TCS, Infosys
3. Explain zero-shot, one-shot and few-shot (N-shot) prompting. ★
4. What is Chain-of-Thought (CoT) prompting? When does it help? ★
5. What is Self-Consistency prompting?
6. What is Generated Knowledge prompting?
7. What is Tree-of-Thoughts (ToT) prompting?
8. What is ReAct prompting?
9. How do temperature, top_p and top_k affect generation? Which settings would you use for a factual Q&A bot vs creative writing? ★ · Asked: TCS, EPAM, Google
10. What are max_tokens and stop sequences?
11. What is frequency penalty vs presence penalty?
12. How do you get reliable structured output (JSON) from an LLM? (Structured outputs, function calling, output parsers.) ★ · Asked: Wipro, Google
13. What is prompt injection? How do you defend against it (including indirect injection via retrieved documents)? ★ · Asked: Wipro, Google
14. How do you version, test and evaluate prompts in production? · Asked: Google
15. What happens when temperature is set to 0? What if it is increased to 1.2? Is temperature 0 fully deterministic? ★ · Asked: Deloitte, Google
16. What is context engineering? How is it different from prompt engineering? ★ · Asked: Google
17. Should you always give more context to an LLM? What is context-window overflow and how do you handle it? ★ · Asked: Google
18. What makes a good enterprise prompt? · Asked: Google