Assignment 13
Fine-tuning LLMs

Part C — Building GenAI Applications

14 questions

★	**Must-know** — asked very frequently in GenAI / ML interviews (6 questions). Prepare these first.
[CODE]	**Hands-on task** — write and run the code (1 tasks).

1. When should you fine-tune an LLM and when should you not? ★ · Asked: Infosys, Google
2. What is Parameter-Efficient Fine-Tuning (PEFT) and why is it needed? ★ · Asked: TCS
3. What are the challenges of full fine-tuning and how does LoRA overcome them? ★
4. Explain how LoRA works. How are the weight matrices updated? (W + BA) ★
5. What are the 'r' (rank) and 'alpha' parameters in LoRA? Which layers do you apply LoRA to?
6. LoRA vs QLoRA — what is the difference and when would you use each? ★
7. What is SFT (supervised fine-tuning)? How do you prepare an instruction dataset and chat template?
8. How do you format prompts for fine-tuning an open-source chat model (e.g. Llama-style [INST] templates)?
9. Which libraries are used to fine-tune LLMs (transformers, PEFT, TRL, bitsandbytes, Unsloth, Axolotl)?
10. What is catastrophic forgetting and how do you avoid it?
11. How do you evaluate a fine-tuned model against the base model?
12. How do you fine-tune Whisper for speech-to-text?
13. Fine-tune a small model with LoRA on a classification or instruction dataset using PEFT + TRL. [CODE]
14. What are the end-to-end steps to fine-tune an LLM for a business use case? ★ · Asked: PwC