Question and Answers-

1)What other Finetuning methods are there apart from SFT-

1. Reinforcement Learning from Human Feedback (RLHF)

- This is a multi-step process that fine-tunes a model based on human preferences.
- It first trains a separate "reward model" to understand what kind of responses humans rate highly, and then uses that model to teach the main language model how to generate more preferable outputs.

---

2. Direct Preference Optimization (DPO)

- DPO is a more recent and often simpler alternative to RLHF.
- It directly optimizes the language model to prefer "chosen" responses over "rejected" ones, skipping the need to train a separate reward model.

---

3. LoRA (Low-Rank Adaptation)

- This is the PEFT method you used in your code.
- It freezes the massive pre-trained model and injects small, trainable "adapter" layers, drastically reducing the memory needed for fine-tuning.

---

4. Instruction Tuning

- This is a broad type of SFT where a model is fine-tuned on a large and diverse collection of tasks formatted as commands or instructions.
- The goal is not to teach it one specific skill, but to make it a better and more general-purpose "instruction-follower."

2) what is the difference between rag and finetuning an LLM-

RAG vs. Fine-Tuning

Fine-Tuning (Changing the Model)

- What it is: Fine-tuning updates the internal weights of the model by training it on a new dataset.
- How it works: You are permanently teaching the model a new skill, style, or a fixed set of knowledge, changing its core behavior.

RAG (Giving the Model Knowledge)

- What it is: Retrieval-Augmented Generation gives the model external knowledge at the time of the query without changing the model itself.
- How it works: It fetches relevant information from a database and adds it to the prompt as context for the model to use. The model's weights remain unchanged.
- Analogy: It's like giving that same student an open book during an exam. They use the book for the answers, but their own memory isn't permanently updated.

Where RAFT Fits In

RAFT (Retrieval-Augmented Fine-Tuning) is a specialized type of fine-tuning.

- Its purpose is to make a base model better at performing RAG.
- You fine-tune the model on data that mimics a RAG setup (like the "golden" and "distractor" documents in your code) to teach it how to better use, cite, and reason about the information provided in the context.
- RAFT is a fine tuning strategy that is practically and efficiently implemented using a method like QLoRA which works by changing a small targeted set of model parameters
- LoRA/QLoRA is a efficient method you use to achieve the goal of making model smarter at using retrieved documents (that is to find the right answer in context and ignore the noise) ,instead of changing all 1.1 billion parameters of the model, QLoRA allows you to change just a tiny fraction of them to teach model a new skill.

3)How can a text message finetune a LLM (Ans:. tokenization)-

A text message fine-tunes a Large Language Model (LLM) by being converted into a sequence of numbers through a process called **tokenization**. The model then learns from the mathematical patterns within these numbers to adjust its internal parameters.

Tokenizer is essentially a translator between human language and world of language model ,it is a tool that breaks down a piece of text into smaller units called taken and convert these token into sequence of corresponding numbers

The process begin with raw text (basically a training example),This text is fed into tokenizer that breaks the text into smaller chunks called tokens, Each unique token is mapped to specific integer ID ,During fine tuning the sequence of numbers is fed into model ,models job is to predict the next number in sequence, It then compares the prediction to actual number ,difference between prediction and reality(error or loss) is used to make tiny adjustments to models internal parameter.