**Why TinyLLaMA + LoRA?**

* **TinyLLaMA** is a small-scale LLM with just 1.1B parameters — making it ideal for local deployment use in our constrained environment.
* **LoRA (Low-Rank Adaptation)** via PEFT enables efficient fine-tuning without retraining the entire model.
  + It reduces memory footprint.
  + It speeds up training.
  + It allows targeted adaptation to domain-specific tasks (like law).

**Steps:**

1. **Environment Setup & Model Loading**
   * Installed necessary libraries (transformers, peft, datasets, accelerate).
   * Loaded the base model (TinyLlama-1.1B-Chat-v1.0) and tokenizer.
2. **Data Preparation**
   * Loading custom legal question-answer pairs.
   * Formated them into prompt-response structure (aligned with ChatML-like input formatting).
   * Tokenizing data with truncation to handle long documents.
3. **PEFT Configuration**
   * Sets up LoRA configuration:
     + r=16, alpha=32 → Balance between compression and performance.
     + Targeted specific transformer modules for injection (q\_proj, v\_proj)
   * Applies the PEFT adapter to the base model.
4. **Trainer Setup**
   * Used transformers. Trainer with custom arguments:
     + Small batch size for limited memory.
     + Gradient\_accumulation\_steps to simulate larger batch.
     + bf16=True and/or fp16=True if mixed precision is supported.
   * Training loop is launched with evaluation steps and checkpoints.
5. **Saving & Merging LoRA Weights**
   * Fine-tuned LoRA weights are saved separately.
   * Optionally, merged with the base model for inference.
6. **Testing Inference**
   * Demonstrates how to load the model + tokenizer and run the chatbot.
   * Includes sample queries related to legal advice (e.g. divorce, maintenance, custody, etc.).

**Why This Was Done**

We are building a **domain-specific legal assistant**. LLMs trained on generic text often perform poorly on niche areas like Muslim Family Law in Pakistan. By fine-tuning TinyLLaMA with **custom legal Q&A**, we:

* Infused domain expertise into the model.
* Enable **offline access** to legal aid using minimal compute.