**Assignment 04 — Distilling a Mini Teacher Model using LoRA**

**Objective:**

This assignment focuses on understanding the concept of model distillation through practical implementation. The goal is to use a large language model to generate high-quality Chain-of-Thought (CoT) reasoning traces and then train a smaller model to replicate this behavior using parameter-efficient fine-tuning techniques.

**Assignment Overview:**

1. **Distillation Goal**

Learn how smaller models can be trained to imitate the reasoning capabilities of larger teacher models, thereby reducing computational costs while preserving quality in inference.

2. **Task Description**

1. Use a large teacher model such as Mistral-7B or DeepSeek-13B to generate CoT answers on a curated set of QA tasks.
2. Fine-tune a smaller model (e.g., DeepSeek-1.3B) using Low-Rank Adaptation (LoRA) or QLoRA on the generated dataset.
3. Perform comparative evaluation between the base model and fine-tuned model.

3. **Workflow Breakdown**

a. **Data Generation**

1. Prepare a prompt-based dataset of reasoning problems.
2. Use the large model to generate CoT-style answers.
3. Save responses in JSON/CSV format with structure:

{

"question": "...",

"cot\_answer": "...",

"final\_answer": "..."

}

b. **Fine-Tuning Setup**

1. Base Model: deepseek-ai/deepseek-coder-1.3b-base
2. Framework: Hugging Face Transformers with PEFT (LoRA)
3. Training approach: Use CPU-safe quantization (e.g., 4-bit QLoRA if available)
4. Epochs: 3–5 (or until convergence on small dataset)
5. Output: Save LoRA adapter in ./student\_model

c**. Evaluation**

1. Compare model behavior on reasoning prompts
2. Measure qualitative improvement in step-by-step reasoning
3. Optional: Track accuracy if dataset has clear correct answers
4. Evaluation Prompts
5. Use a consistent set of test questions (5–10 samples) for comparing outputs from:

* Base model (before fine-tuning)
* Student model (after distillation)

**Sample Comparison Prompt:**

Q: A rectangle is 5 meters long and 3 meters wide. What is its area?

A: [Expect student model to respond with step-by-step reasoning learned from the teacher]

**Software Requirements**

1. Python ≥ 3.10
2. Transformers ≥ 4.40
3. bitsandbytes (if using GPU quantization)
4. peft
5. accelerate
6. torch ≥ 2.1

**Installation:**

pip install transformers peft accelerate bitsandbytes

Training Configuration Tips

Use CPU mode (float32) if no GPU is available

Ensure reproducibility with seed and checkpoint saving

Save evaluation logs and generated outputs for comparison

**Deliverables**

student\_model (LoRA adapters)

**Author:**

Bhagyesh Makhamale

Under guidance of: Dr. Yogita Mahatekar

Affiliation: COEP Technological University — Department of AI & Data Science

Academic Year: 2025–2026