Task Specific Adaptation of Small Language Models

Sanjay Shrestha (THA079BCT039)

Deep Shrestha (THA079BCT000)

Rohan Dhakal (THA079BCT000)

Pradeep Pokhrel (THA079BCT000)

December 2025

INTRODUCTION

1.1 Background

Advent of LLM have fundamentally changed the software development and code writing process. LLM have become integral part of the workflow for developers and students. Even with this much of success, LLMs rely on massive datasets and big GPUs for training and running these models which makes them impossible for student and general people to run and train locally.

Fine-tuning is the process of further training a pre-trained LLM on a smaller, task-specific dataset. While the initial pre-training gives universal linguistic knowledge, fine-tuning shapes this generalized competence into specialized expertise.

The OPT-350M (Open Pre-trained Transformer) model, developed by Meta AI, is a decoder-only LLM. With its 350 million parameters, it gives an important balance. It is large enough to possess meaningful generative capacity, yet small enough to be computationally efficient for research, development, and fine-tuning on consumer-grade or limited-resource hardware. This makes it an ideal candidate for demonstrating efficient specialization techniques.

1.2 Objectives

A. Making better code generator than existing OPT-350M by finetuning it on python code datasets.
B. Using PEFT through QLoRA to finetune base model i.e OPT-350M.