Project Overview

- **Project Title**：Generating Personalized Study Plan with Smaller Language Model
- **Team Members**：Christoph Winter, Ying Zeng

Project Milestons

- **Period1: Model Finetuning**
  - **Target**: Improve model's domain knowledge on math
  - **Checkpoints**
    - [x] Build training and test pipeline
    - [x] Finetune Llama 3B/8B model with GSM8K dataset using LoRA
    - [ ] [WIP] Compare accuracy/solveRate with unfinetuned pretrained models
- **Period2: Prompt Engineering**
  - **Target**: Find the best prompt to generate the most personalized study plan
  - **Checkpoints**
    - [ ] Define a metric to evaluate the quality of a study plan by splitting the score into sub-aspects (e.g. fluency, the level of personalization, etc.)
    - [x] Build inference framework
    - [ ] Explore the best approach to craft prompts: using only prompt text or prompt tuning.
- **Period3: Mobile Phone Deployment**
  - **Target**: Quantize the model to be small enough for depolyment on a phone while maintaining a certain level of performance
  - **Checkpoints**
    - [ ] Model Quantization and Deployment
    - [ ] Model inference on a mobile phone

Obtained Milestones

- Build training/test/inference pipeline using Unsloth => Github
- Get the first version of the fine-tuned model
- Compare accuracy with unfinetuned pretrained models

  => Blocking in testing (consuming more time than expected)

Bottlenecks

- **For Model Finetuning**
  - Model inference is not as efficient as training taking more than 2hrs for testing while training only takes 15mins

    => need to debug or look for more efficient libraries

    - Model perfomance might not overcome existed LLMs

      => might need to finetuned on multiple math word problem datasets

- **For Prompt Engineering**
    - How to score the qulities of different study plans

      => might need several experiments

- **For Model Deployment**
    - How to merge the best prompt in the backend so that users can achieve the best model performance with minimal input

Work Contribution

- Christoph Winter
    - Majority Deployment Implementation + Partial Finetuning Implementation
    - Prompt Engineering Design and Implementation
- Ying Zeng
    - Majority Finetuning Implementation + Partial Deployment Implementation
    - Prompt Engineering Design and Implementation