# Task 1: Fine-tune LLM for Language Translation

Saptarshi Bhattacharya
RPTU , Department of Computer Science

***Note:*** *This report contains a project documentation and reflection on the portfolio task submitted for the lecture Engineering with Generative AI in WiSe 2024-25. This report is an original work and will be scrutinised for plagiarism and potential LLM use.*

## 1 Portfolio documentation

Compile a comprehensive documentation of your project, including all the project phases. You will need to explain every choice you made during the project and your thoughts about the results you get. You will introduce the results in suitable visualisation. Furthermore, you will need to explain which criteria you follow to build your prompts and how they affect the results.

Students write the entire documentation with sections, sub-sections, diagrams, etc in this section. Please write as comprehensively as possible. Head to the document 1_documentation.tex. You are free to use as many subsections as required. We will not provide a template for documentation.

### 1.1 Task Description

The goal of this project was to fine tune a general purpose Large Language model for German to french language translation using synthetic datasets, and to evaluate if a synthetic dataset provides any benefits over using a benchmark dataset. This project was developed in three phases, research, design, and implementation.

### 1.2 Research

#### 1.2.1 Dataset

Our goal was to find a dataset with pairs of inputs for the German-French lanugage translation. I found the Flores-200 dataset from facebook as the ideal candidate, because of the following reasons.

1. It is a standard benchmark dataset that was developed by facebook and is refered by many papers for NLP and translation tasks.
2. It has sufficient amount of direct german to french sentence pairs for our task.
3. It is available on huggingface, which made it convenient to use.
4. The dataset was made into 80-20 train test split, as the number of sentence pairs chosen was restricted to 1000, hence, I wanted sufficient number of sentence pairs available for testing, hence a split such as 90-10 was rejected.

### 1.2.2 Model

Here, the goal was to select a relatively small model which was not less than 1B parameters in size. The model was required to be a pre-trained general-purpose LLM in the first step and not a language translation model or a model with random weights. Here, I found Qwen2-1.5B to be the best option because of the following reasons

1. It has only 1.5B parameters, which makes it suitable to be fine tuned within the resource bounds of google colab free tier.
2. Although it is a general purpose LLM, it performs well for multi-lingual tasks.
3. Bigger models such as R1-llama 8B were causing memory issues, and models of similar size were often generating gibberish, which was not an issue with Qwen.

## 1.3 Design

### 1.3.1 Tuning Approach

I chose LoRA (Low-Rank Adaptation) for fine-tuning as full fine-tuning was not feasible with the given resources. It is a parameter-efficient fine-tuning method that reduces the number of trainable parameters by freezing most parameters and injecting low-rank matrices into the model. This makes it computationally efficient and hence, well-suited for fine-tuning Qwen within the limited resources of google Colab.

### 1.3.2 Prompt Design

I created the prompt with the aim of producing synthetic data using a more advanced model, specifically Qwen/LLaMA3-70B-8192. Here's my approach:

1. **Creativity and Uniqueness**:
    a) **Defining the Task**: I clearly specified the objective of generating new translation pairs from German to French.
    b) **Ensuring Variety and Precision**: I prioritized diversity, accuracy, and natural flow in the sentences to make them realistic and engaging.
    c) **Organized Structure**: I implemented a fixed format for the output to ensure uniformity and practicality.
2. **Functionality**:
    a) **Guiding Examples**: I incorporated sample translations from the dataset to help the model produce relevant and high-quality outputs.
    b) **Detailed Guidance**: I included thorough instructions to minimize repetition, formatting mistakes, or incomplete results.

### 1.3.3 Evaluation Metric Selection

**BLEU Score**: The BLEU (Bilingual Evaluation Understudy) metric was selected to assess translation quality due to the following reasons:

1. **Established Standard**: BLEU is a commonly used metric in machine translation research, allowing for consistent comparisons across different models and studies.
2. **Focus on Precision and Conciseness**: BLEU evaluates how well the generated translation aligns with reference translations by analyzing word overlap while considering both accuracy and brevity.
3. **Objective Performance Measurement**: BLEU assigns a numerical score to translations, facilitating direct performance comparisons between models.

Given its reliability and widespread recognition, BLEU serves as a suitable metric for evaluating translation quality in this project.

## 1.4 Implementation

Some of the finer implementation details are as follows:

1. The code for translation and testing was put into a function so that it could be reused easily.
2. Similarly the model fine-tuning was also done via a function for similar reasons.
3. Models were deleted from memory as soon as they were evaluated so that memory is available for other models to enter. Cache was also cleared before every fine-tuning step.
4. Groq was used to access the larger LLM via an API, as it is not possible to load it directly in Google Colab.

## 1.5 Testing and Evaluation

1. **BLEU Scores for the Models**
   a) **Model A(Base)**: 0.210
   b) **Model B (Fine-Tuned on Dataset A(Train))**: 0.211 - Fine-tuning with Dataset A did not significantly improve the BLEU score, indicating that the model retained similar performance.
   c) **Model C (Fine-Tuned on Synthetic Data)**: 0.211
   d) **Model D (Fine-Tuned on Combined Data)**: 0.208
2. **Analysis and Insights**
   a) The minimal improvement in BLEU scores across fine-tuned models suggests that the base model was already well-trained for German-to-French translation.
   b) Fine-tuning on real or synthetic data alone yielded nearly identical results, indicating that the model's prior knowledge was strong enough that additional training had little impact.
   c) The slight decrease in Model D's BLEU score may suggest that mixing real and synthetic data introduced inconsistencies, affecting performance.

## 2 Reflection

**In 3-5 pages, 1500-2000 words**

This section needs to be adjusted to align with the reflection requirements specified in the selected task.

**Note:** You should address all the questions from your selected task. Please list each question and provide your answers in the following enumeration.

1. **What was the most interesting thing that you learned while working on the portfolio? What aspects did you find interesting or surprising?**

   **Answer:** The most interesting thing I learned while working on the portfolio was that relatively smaller LLMs can be fine-tuned for specialized tasks with impressive performance. Before the portfolio, I believed that only larger LLMs could perform such tasks with higher accuracy. However, Qwen was able to provide consistent responses as well.

   I was also surprised that there was no performance increase between model_C and model_D. Since I observed improvements from model_A to model_B and then from model_B to model_C, I expected a similar trend from model_C to model_D. Instead, I encountered a performance decrease, which was unexpected.

2. **Which part of the portfolio are you most proud of? Why? What challenges did you face, and how did you overcome them?**

   **Answer:** The part of the portfolio I am most proud of is the fine-tuning process using LoRA (Low-Rank Adaptation). This was a challenging yet rewarding aspect of the project. Fine-tuning a large language model like Qwen required careful consideration of computational resources, hyperparameter tuning, and dataset preparation. Successfully implementing LoRA allowed me to adapt the model efficiently without requiring excessive computational power.

   One of the main challenges I faced was memory management. Fine-tuning large models is resource-intensive, and I initially struggled with out-of-memory errors. To overcome this, I implemented 4-bit quantization, which significantly reduced the memory footprint of the model. Additionally, deleting models from memory after evaluation helped manage resource constraints.

3. **What adjustments to your design and implementation were necessary during the implementation phase? What would you change or do differently if you had to do the portfolio task a second time? What would be potential areas for future improvement?**

   **Answer:** During the implementation phase, several adjustments were necessary to improve the model's performance and efficiency. For example, I initially used a larger batch size for training, but this led to memory issues. I reduced the batch size and increased the gradient accumulation steps to maintain training stability. Additionally, I experimented with different learning rates and found that a lower learning rate (2e-4) worked best for fine-tuning.

4. **Include a brief section on ethical considerations when using these models for language translation tasks.**

   **Answer:**

   a) **Bias and Fairness:** Language models may exhibit biases, such as gender stereotypes or cultural biases, which can lead to unfair or inaccurate translations. For example, a model might default to male pronouns in certain contexts, reinforcing gender stereotypes.

   b) **Privacy and Data Provenance:** Sensitive data used in training could expose private information, such as personal names or confidential details.

   c) **Misuse Potential:** Language models can be exploited for malicious purposes, such as generating disinformation or circumventing content moderation systems.

   d) **Transparency and Accountability:** Users may place excessive trust in machine-generated translations, especially in critical areas like law or medicine.

   e) **Cultural Sensitivity:** Direct translations might miss cultural nuances, leading to misunderstandings or offense.

5. **From the lecture/course, including guest lectures, what topic excited you the most? Why? What would you like to learn more about and why?**

   **Answer:** The Amazon guest lecture, where the speaker explained how generative AI is utilized within Amazon, was the most exciting. I was able to understand how the industry approaches cutting-edge research topics such as generative AI. It was fascinating to see how customer perspectives are considered, even for technologies that lack well-defined products.

   In the future, I would like to learn more about integrating generative AI with existing technologies without significant performance penalties. For example, the generative ad shown in the lecture demonstrated how simple product photos could be transformed into compelling advertisements.

6. **How did you find working with the DIFY platform during the coursework? Would you recommend using DIFY for learning Generative AI technologies, and why? What is the best way to start learning Generative AI—through Python code or no-code platforms—and why?**

   **Answer:** I would recommend using DIFY for learning generative AI, especially for beginners. Its no-code approach allows users to focus on understanding the concepts and applications of generative AI without being overwhelmed by technical details. However, for those who want to dive deeper into the underlying mechanisms, learning through Python is essential. Python provides greater flexibility and control, enabling users to customize models and implement advanced techniques.

   I found working with DIFY enjoyable, but from a reusability standpoint, it was challenging. Additionally, managing APIs through DIFY was cumbersome. While DIFY is an excellent design tool, I believe Python is still superior in terms of support and customizability.

7. **How did you find the assignments and exercises in the course, and how did they help you in the portfolio exam?**

**Answer:** The assignments and exercises were highly informative. I believe that any student who completes the assignments and exercises should be able to perform well in the portfolio exam. The exam closely mirrors the topics covered in the assignments and exercises. Throughout the portfolio exam, I frequently referred to my solutions from these exercises.

## References