Natural Language Processing (CO3086)
NLP 242: ASSIGNMENT

HO CHI MINH UNIVERSITY OF TECHNOLOGY
Vietnam National University Ho Chi Minh

1 Team Selection

Students are required to form groups of 4 to 5 members and register their groups using the following link (Group Registration Link). The deadline for group registration is March 27, 2025.

2 Task Description

In 2017, the introduction of the Transformer model brought significant improvements to various tasks in the field of Natural Language Processing (NLP). The Transformer architecture, introduced in the paper "Attention Is All You Need" by Vaswani et al., replaced traditional recurrent and convolutional neural networks with self-attention mechanisms. This innovation led to more efficient parallel processing, improving both training speed and model performance. Transformers became the foundation of many state-of-the-art NLP models such as BERT, GPT, and T5.

Since then, language models have grown exponentially in size, with an increasing number of parameters and larger training datasets. Notable advancements include OpenAI's GPT series (GPT-3, GPT-4) and Google's PaLM models, which contain billions to trillions of parameters. These models have demonstrated remarkable capabilities in text generation, machine translation, and various NLP applications.

To efficiently utilize large language models (LLMs), fine-tuning techniques have been developed to adapt pre-trained models to specific tasks or new datasets. However, full fine-tuning is computationally expensive due to the massive size of these models. As a result, various parameter-efficient fine-tuning (PEFT) methods have been introduced to reduce the computational cost while maintaining high performance.

3 Submission Instructions

Students are required to write a report on the development of Transformer models, their increasing size, and different fine-tuning techniques. Additionally, students must conduct experiments by selecting a few large language models and applying at least three fine-tuning techniques to improve performance on at least three different tasks.The report should include:

- An overview of Transformer models and their impact on NLP.
- A discussion on the evolution of large language models, including key milestones and examples.
- A detailed explanation of various fine-tuning techniques and their advantages/disadvantages.
- Experimental results comparing different fine-tuning methods on multiple tasks.
- Insights and conclusions based on the experiments conducted.

Final submissions should include both the written report and code implementations for the experiments. The deadline for assignment submission is June 1, 2025.