Project 1 - Fine-Tuning LLMs
Trustworthy ML (ECE 696B): Spring 2025

- Due Date: **Sunday, Feb 23, 2025 by Midnight Tucson Time**
- Mode of submission: Submit a 2-3 page report on D2L
- Maximum Credit: 200 points

Project Goals and Objectives– the goals of this project are three fold:

- Understand how to set up, run, and evaluate open-source LLMs.
- Gain hands-on experience fine-tuning models for improved performance.
- Learn best practices in experiment tracking, reporting, and evaluating LLMs.

You will explore one or more tasks (e.g., classification, language modeling, summarization) and document your findings. If you "borrowed" and adapted code from some open source website, you must properly cite that in the report. You should a typed report in PDF format (using LaTeX) with your results and discussions as well as link to the code repository.

Slack Channel for Discussion. Discussion between students is strongly encouraged. However, the code, implementation, and reports must be created by each student individually. I am creating a Slack channel for discussions. You can sign up using the link below.

Slack Link for Course Discussions: https://tinyurl.com/c69n2ruh

1 Project Tasks and Requirements

1.1 Datasets and Models

We will be working with **small open-source LLMs** (e.g., GPT-2, GPT-Neo, DistilGPT-2) and at least one dataset from the following options:

- **Classification:** IMDB (50K movie reviews) or AG News (127.6K news articles)
- **Language Modeling:** WikiText-2 (under 1M tokens)
- **Summarization:** SAMSum or a small subset of CNN/DailyMail

1.2 Task 1: Baseline Evaluation

- **Download and set up** at least two open-source, small LLMs (e.g., GPT-2 Small, GPT-Neo 125M).
- **Select a dataset** and split it into train/validation/test sets.
- **Run inference** to measure baseline performance:
  - For classification: measure accuracy, F1-score, or another relevant metric.
  - For language modeling: measure perplexity (on WikiText-2 or other).
  - For summarization: measure ROUGE scores (if choosing a summarization dataset).
- **Record and analyze** results. Provide sample model outputs to illustrate baseline behavior.

1.3 Task 2: Fine-Tuning and Improved Evaluation

- **Fine-tune** the same models from Task 1 on your chosen dataset(s).
  - You may choose full fine-tuning or a parameter-efficient approach (e.g., LoRA).
  - You should use UA HPC resources (available for free) for this project.
- **Evaluate** the fine-tuned model(s):
  - Compare the new results (accuracy, perplexity, ROUGE, etc.) to the baseline.
  - Provide at least one performance table or chart showing improvement.
- **Discuss observations**:
  - How much improvement did you see?
  - Were there any unexpected results?

2 Deliverables

Your final submission should include:

1. **Project Report (submitted to D2L):**
   - Description of the dataset(s), models, and methods used.
   - Results (tables, charts) and discussion of performance.
   - Example model outputs (e.g., classifications, summaries, or generated text).
2. **Source Code:**
   - Well-commented scripts or notebooks for dataset processing, model training, and evaluation.
   - A `requirements.txt` (or `environment.yml`) for dependencies.