Course Project: LLM or Not?

1 Project Background

It's getting harder and harder to tell whether a piece of writing came from a person or a machine. Your job in this project is to build a **binary classifier** that looks at an essay and decides: was this written by a student, or did an LLM write it?

2 Dataset Description

You are provided a single file, `data.csv`, containing essays along with their labels. Each essay was written in response to one of several prompts that asked the writer to read some source material and then compose a response. Some of these essays are genuine student work; others were produced by various LLMs.

Column	Type	Description
`id`	string	Unique identifier for each essay
`text`	string	The full essay text
`label`	string	`positive` = LLM-generated, `negative` = student-written

Table 1. `data.csv` columns

Important: We will evaluate your model on a held-out test set that is *not* included in the release. Do not overfit to the provided data.

If you borrow or adapt code from a public sources, you must cite it in your report.

3 Objective

For each essay in the held-out test set, output the **probability** that it was generated by an LLM. Submissions are scored by **ROC-AUC**, so submit probabilities (floats between 0 and 1), not hard labels.

4 Deliverables

The following are the deliverables of this project:

- A binary classifier to label each essay as `positive` (LLM-generated) or `negative` (student-written).
- The number of parameters in the model must not exceed 15 million.
- You are free to use any publicly available model (pre-trained or otherwise) with or without augmentation, but it is not a requirement. You can also augment the data as you see fit.
- You may not incorporate any additional data from external sources. Train only on what is provided.
- You should be constructing the features and labels from the given data for training purposes, as well as the training and testing pipeline.

5 Submission

1. Submit all your code, including training and evaluation, as a `.zip` file.
2. Submit `prediction.csv` file for scoring and evaluation.
3. Submit a 2-page report (1-inch margin, 12-point font) and include
   - Your approach, model, and any other design choice.
   - Hyperparameters that you used for training.
   - Training and test results.
   - Any other interesting details about the approach or model.

Submit your predictions to the class Kaggle competition. Your file must be a CSV with exactly two columns: `Id` and `Prob`. For each `Id` in the test set, predict the probability that the essay is LLM-generated (`positive`). Use the original IDs from `test.tsv`; do not renumber them. The file should look like this:

```
Id,Prob
23aa,0.8
9f1b,0.9
c04e,0.3
```

Column names are case-sensitive and must match exactly. To submit, click *Submit Prediction* in the upper-right corner of the Kaggle competition page and upload your CSV. Include the names of all group members in the submission description. After submitting, go to the *Submissions* tab and click *Select* next to your best submission so it appears on the leaderboard.