Project: Fine-tune a Small Language Model (SLM) for Summarization

Your task is to fine-tune a Small Language Model (SLM) (smaller than 7B, e.g., Qwen2.5-0.5B-Instruct, Atlas-Chat-2B, mT5-Base) on a summarization task. You will use Google Colab for this purpose.

Protocol:

1- Dataset Selection: Choose / prepare an Arabic-language (e.g., MSA, Darija), unannotated dataset (e.g., Arabic Wikipedia, news articles, etc..) containing 5,000 documents.

2- Dataset Annotation: Create synthetic summaries of the 5,000 documents using Large Language Models (LLMs, e.g., Atlas-Chat-27B, Jais-adapted-13b-chat, Qwen2.5-32B-Instruct) of ***high quality***.
   Ensure the model can run on a Google Colab *free-tier* GPU. If necessary, you can apply quantization techniques like AWQ to fit the model within Colab's hardware constraints.

3- Data Splitting: Split the annotated dataset into training and test sets (including a validation set is optional but encouraged).

4- Model Finetuning: Fine-tune the SLM(s) on the training set.

5- Model Evaluation: Evaluate performance on the test set, with measures such as ROUGE, BERTScore, LLM-as-a-Judge.

Notes:

- ***Justify your choices and decisions at each step.***
- Any improvements or optimizations to the proposed protocol are highly appreciated.
- Feel free to include any additional steps or considerations, such as hyperparameter tuning or more sophisticated evaluation methods.

Rules:

- You are given three weeks (until 21/02/2025 end of the day).
- Each team must be composed of up to 2 students.
- You will present your solution in a 30-minute oral session.
- The course grading is distributed as follows: 10% for in-class quizzes, 40% for the oral assessment, 40% for the report, and 10% for code quality.

Submission:

Please submit a zipped folder named Lastname_Firstname.zip (of the submitter); Each team only needs to submit one zip file by one of the team members. Please ensure that the real names of all team members appear on the cover page.

https://docs.google.com/forms/d/e/1FAIpQLSfcI7dvEORAIRCRZFNsjszBCL19-Kj837uC2ns4DukAlAtkdA/viewform?usp=dialog

The zipped folder should include:

- A “code” sub-folder containing all the scripts needed to reproduce your submission.
- A “data” sub-folder containing all the data needed to reproduce your submission.
- A PDF report of max 8 pages for the main content, excluding the cover page and references. In addition, you can use extra pages of the appendix (for prompts, explanations, algorithms, figures, tables, link to the online data resources, etc.).

Contact: guokan.shang@mbzuai.ac.ae