Gen AI with LLMs: Individual Assignment

1. What is required for a grade of B on the Assignment

A B-level Individual Assignment meets all of the following criteria at a “meets expectations” level, with no major correctness or reproducibility issues.

Going above and beyond in any of the areas merits a grade better than B.

2. Individual Assignment task summary

2.1 Training process

- Fine-tune an **open-weight** LLM on a clearly defined task (e.g., classification, QA, or instruction-following) using a coherent dataset, and produce a checkpoint or adapter

- Provide runnable code or notebooks that cover data loading, training, checkpoint saving, and basic inference, with at least 5 hyperparameter tuning configs tested and a brief justification of the fine-tuning method used (e.g., LoRA/QLoRA vs full FT) as well as on the choice of final hyperparameters
- Monitor at least 1 training signal (loss curve, validation metric) to justify stopping training

2.2 Evaluation

- Define at least one appropriate quantitative metric for the task (e.g., accuracy/F1, exact match, or a simple rubric-based rating for generations)

- Compare base vs fine-tuned model on a small but non-trivial evaluation set (around 20–30 examples or a held-out split), and present results in a simple table or figure

2.3 Analysis and reporting

- Submit a short report or notebook with markdown text that explains your task, dataset, model choice, fine-tuning process, evaluation, and results.

- Provide a brief analysis of where the fine-tuned model improved (or did not), including at least one observed limitation or failure mode, and mention at least one alternative design choice if the model did not improve after fine-tuning.

2.4 Reproducibility and ethical use

- Include enough implementation details (model name, dataset sources, key hyperparameters, environment notes) for another student to reproduce the experiment, if not all the above are contained in a notebook

- Use models and data with appropriate licenses, cite sources, and mention at least one relevant ethical or risk consideration (e.g., bias, harmful outputs) for the chosen task and how it could be addressed