WHY: To reduce environmental impact of LLM and highlight sustainable LLM lifecycle

WEEK NO.	WHAT	TASKS	SUB-TASKS	COMMENTS	ESTIMATE
1	Starting the project with a particular task - creating a model to estimate whether a review is good or bad and we want to know the carbon emissions and performance upon training and testing of a standard fine-tuned BERT. This provides us with a baseline benchmark which we are going to improve upon.	Dataset & Environment Setup	• Download IMDb dataset • Split into training and testing • Tokenize to feed into transformer		1 Day
		Fine-tune BERT (for sentiment analysis)	• Load BERT-base-uncensored model • Set hyperparams like no. of epochs, learning rate, batch size • Train and evaluate using performance metrics like precision, recall, accuracy and F1 score and make confusion matrix • Hyperparameter tuning		2 Days
		Carbon Tracking & Metrics	• Integrate CodeCarbon for tracking training carbon emissions data • log all data and metrics		2 Days
		Inference on local machine	Provide different prompts and track carbon emission for inference		1 Day
2	Apply model pruning optimization technique on fine-tuned BERT and verify whether it fares better than the standard model. This shall provides us with proof that model optimization does reduce carbon footprint as well. However, we also keep in mind the performance of the optimized model so as to compare the reduction or increase in performance measure.	Learn how to implement model pruning	• Code a tutorial model pruning to familiarize with it • Learn how to implement model pruning on a fine-tuned model		2 days
		Implement Model Pruning on fine-tuned BERT	• Apply both unstructured and structured pruning • Evaluate using performance metrics		1 day
		Metrics Logging	• Integrate CodeCarbon and measure carbon emission data • Log performance and carbon emissions data		1 day
		Inference on local machine	Provide different prompts and track carbon emissions on inference		1 Day
3	This is the part where Cloud comes into the picture. We move our models from local testing to the cloud deployment phase. Now we actually try verify our hypothesis that model optimization as well as sustainable cloud configuration can work together to reduce the carbon emissions. We also see how much of a hit accuracy and other performance metrics might take.	Dockerization	Create a Docker file for both standard and pruned model as well as a basic web UI to work as an inference		1 day
		Cloud Run Deployment	• Build Docker image for both models • Upload to Artifacts registry • Deploy model on Cloud Run		1 day
		Measure Metrics	• Go to Cloud Run service to get metrics like cold start latency, response time • Compare energy estimates using GCP's Carbon Footprint tool		1 day

4	This is where we visualize the whole picture and see the actual numbers. We compare how the both model fare against each other in terms of carbon emissions and performance of training, inference and cloud deployment and maintainence.	Analyze Results	Compile all performance metrics and carbon emissions data		1 day
		Visualize Trade-offs	Create graphs/tables: accuracy vs. energy, etc.		1 day
		Write MVP Report	Summary of findings, insights		1 day