

Assignment 4

1. Fine-tune answerdotai/ModernBERT-large on the clinc_oos (load_dataset("clinc_oos", "plus")) dataset. Show the main training results and show the result from giving the model the following query: “Hey, I'd like to rent a vehicle from Nov 1st to Nov 15th in Paris and I need a 15 passenger van”.
2. Follow the example from Chapter 8 in Natural Language Processing with Transformers and produce a plot where you compare the accuracy, latency and size of 1) your fine-tuned ModernBert-large baseline, 2) a distilled version, 3) a distilled version using Optuna and 4) a quantized-distilled version. Think carefully about the appropriate student model to select.

Attach your Jupyter notebook as an appendix. Include only code relevant to solving the problems.