| **Model** | **WER** | **CER** |
|---|---|---|
| Base Model (Task 2) | 14.39% | 6.06% |
| Finetuned Model (Task 3) | 8.11% | 3.36% |

**Key Observations:**

1) Substantial Accuracy Gain:

Fine-tuning on the Common Voice data yields a noticeable reduction in word and character error rates, indicating the model adapts well to the domain of the CV-valid-dev set.

2) Consistent Performance Gains:

CER is consistently lower than WER, indicating that most errors occur at the word level (e.g., substitutions of whole words) rather than character-level mistakes.

**Proposed Steps to Further Improve Accuracy:**

1) An expected further improvement of model performance would come from utilizing the full training dataset if we have sufficient RAM resources. Alternatively, a lower batch count during training and higher gradient_accumulation_steps can compensate for the RAM usage with more training time.
2) Data augmentation could be performed on the audio data, e.g. through low pass filtering, varying the tempo, or introducing artifacts like echo to artificially increase the training dataset. Alternatively, can also introduce time or frequency masking to improve robustness to real-world variations.
3) The additional metadeta that came with the datasets: age, gender, accent could be incorporated into the model by adding additional layers on top of the base model for the extra inputs. These information can help the model make better predictions.
4) Learning Rate Schedules & Optimizers: Experiment with different learning rates, warmup steps, or optimizers (AdamW vs. Adam) to find the best training stability and convergence.