After completing the task, I can suggest a few improvements. The most critical is enhancing the quality and size of the dataset. A larger dataset typically leads to better model performance, as it allows the model to learn from a more diverse set of examples.

Another issue was the quality of the dataset itself. Using ChatGPT to generate samples was helpful, it introduced some problems, particularly with token labeling, where inconsistencies and inaccuracies were common. This means that the generated data requires thorough manual review and correction to ensure labeling accuracy.

Moreover, it would be beneficial to experiment with different hyperparameters. Fine-tuning parameters such as learning rate, batch size, and number of epochs can significantly impact model performance. Running multiple experiments with varying configurations will help find the optimal setup for the best results.