Task 4
i)

Model Type	Training Loss	Test Set Performance
Odd Layer	0.1592	0.322
Even Layer	0.1604	0.317
LoRA	0.0561	0.096

The LoRA model outperformed both the distillation based approaches by achieving the lowest training loss and highest test accuracy. Among the distilled models, the Even Layer model performed slightly edged out the Odd Layer model in test accuracy (97.84% vs. 97.72%), though it exhibited higher divergence loss and more variance in training. In contrast, the Odd Layer model showed the most stable training behaviour and the lowest classification loss among the distillation based variants.

The LoRA based model has been trained on the same dataset with the same seed values and with the same learning rate as the 6 layer student models, i.e. both odd and even, and it exhibits a better accuracy and lower training loss in the evaluation metrics as compared to the other two 6 layered distillation based models. But during testing for multiple sentences on all three models, LoRA based model was not able to perform well and was not able to produce ⅓ rd of the results as produced by the distillation based student models. Some derogatory words were specifically given in as the input but the LoRA based model was not able to mark them as toxic while the other two models easily labelled those words as toxic. Same happened when those words were placed into a sentence and was passed to the models, the output achieved was almost similar to the one just discussed.

ii) During the implementation of the student models (even and odd layered) and LoRA based model, several challenges were encountered:-
While LoRA showed superior performance on quantitative metrics (accuracy, loss), its predictions were mostly less interpretable and less aligned with human intuition when compared to distillation based models.
LoRA simplifies fine-tuning by updating only lightweight adapters within the full 12-layer architecture, it inherently requires running computations through all layers, leading to higher computational overhead and longer training times compared to the distilled models

We can combine distillation with LoRA which might leverage the benefit of both approaches.