For the final solution, I have fine-tuned and tested several models from Hugging Face Seq2SeqTextGeneration models such Bert, T5, Blenderbot, and MVP, and compared them with RadarPlot.

Problem I faced is that I couldn’t train all the model at once, so I just saved the test results, and visualised the metrics of each model to choose the best one.

I have used the original dataset from para-NMT-detox and, in addition, I have usedtoxic comments from Wikipedia datasetfor testing the network.

Next,I have fine-tuned and tested several models from Hugging Face Seq2SeqTextGeneration models such Bert, T5, and MVP, which gave the nearly similar performance for every metric. The results were promising with a BLEU score of almost 30.

Finally, I have made radar plot with 3 metrics: BLEU score, brevity penalty, and average inference time, where each model's test results is compared (1000 values from dataset).

