Reflection & Ethics

The challenges I faced during solving:

I faced several technical and engineering challenges while working. First, in model training (fine_tuning.ipynb), loading the data from all.json required careful text processing, especially with mature jokes that could be long or inconsistent, which led to tokenization issues and exceeding the maximum length (max_length=128). Training also took a long time on the GPU, with memory errors occurring when increasing the batch size. In text generation (text_generation.ipynb), the results were sometimes inconsistent, such as generating repetitive or illogical stories due to the temperature setting (temperature=1). As for summarization and question answering (summarization_qa.ipynb), the summarization was sometimes inaccurate, especially with sensitive articles like article.txt about Gaza, where the model ignored important details such as 'hunger' and gave short, incomplete answers. I also encountered library compatibility issues, such as with transformers in the local environment.

Biases and ethical concerns:

I noticed clear biases in the data and models. In all.json, some jokes contain racial or gender biases (such as jokes about "Jews" or "women"), which reinforces stereotypes if the model is trained on them. The GPT-2 model, trained on extensive internet data, tends to generate sexually or culturally biased content, such as in stories that diminish the role of women. Ethical concerns include the spread of misinformation in summarization, especially on sensitive political topics like Gaza, which can mislead readers or reinforce biased narratives. There are also concerns about privacy (use of Reddit data) and exploitation, as the model could be used to produce offensive or harmful content.

Recommendations for safe use:

To ensure safe use, I recommend pre-checking the data to remove biases and using techniques like debiasing during training. Ethical approvals should be obtained before collecting data, and human monitoring of outputs should be done to avoid harmful content. For sensitive topics, use specialized models like bart-large-cnn with manual verification. It is also preferable to train on culturally diverse data and to apply rules such as not generating violent or politically controversial content. Finally, raise awareness among users about the risks of artificial intelligence.