**Meta Llama 3.1 8B**

**Dataset:**

We talked to chatGPT 4o and 4o mini and then exported the chats in a JSON format to compile 700+ Question-Answer pairs.

**Problems:**

* *Memory Constraints:*

Loading the full 32-bit model is not an option as the available memory is less. Can not load 8- and 16-bit quantized versions. The problem with 8 bit is that the GPUs aren’t being set up in parallel by Kaggle, while 16bit exceeds the memory limit.

* *Results Before Fine-Tuning:*

The results of the full 32-bit model, which were tested using Groq’s API key, were good enough to be considered, however a robotic behavior was observed.

The 8-bit model was not able to communicate in Roman Urdu therefore we had to fine-tune it on the Question Answer Dataset.

The 4-bit model produced even worse results.

* *Results After Fine-Tuning:*

We were only able to fine-tune the 4-bit model. The results improved only with respect to language, but the basic knowledge was lost as the model was 4-bit quantized.

**Inference Time:**

The inference times were very similar for the versions that we could test, however the Groq version produced the results in under a second due to it’s technology.