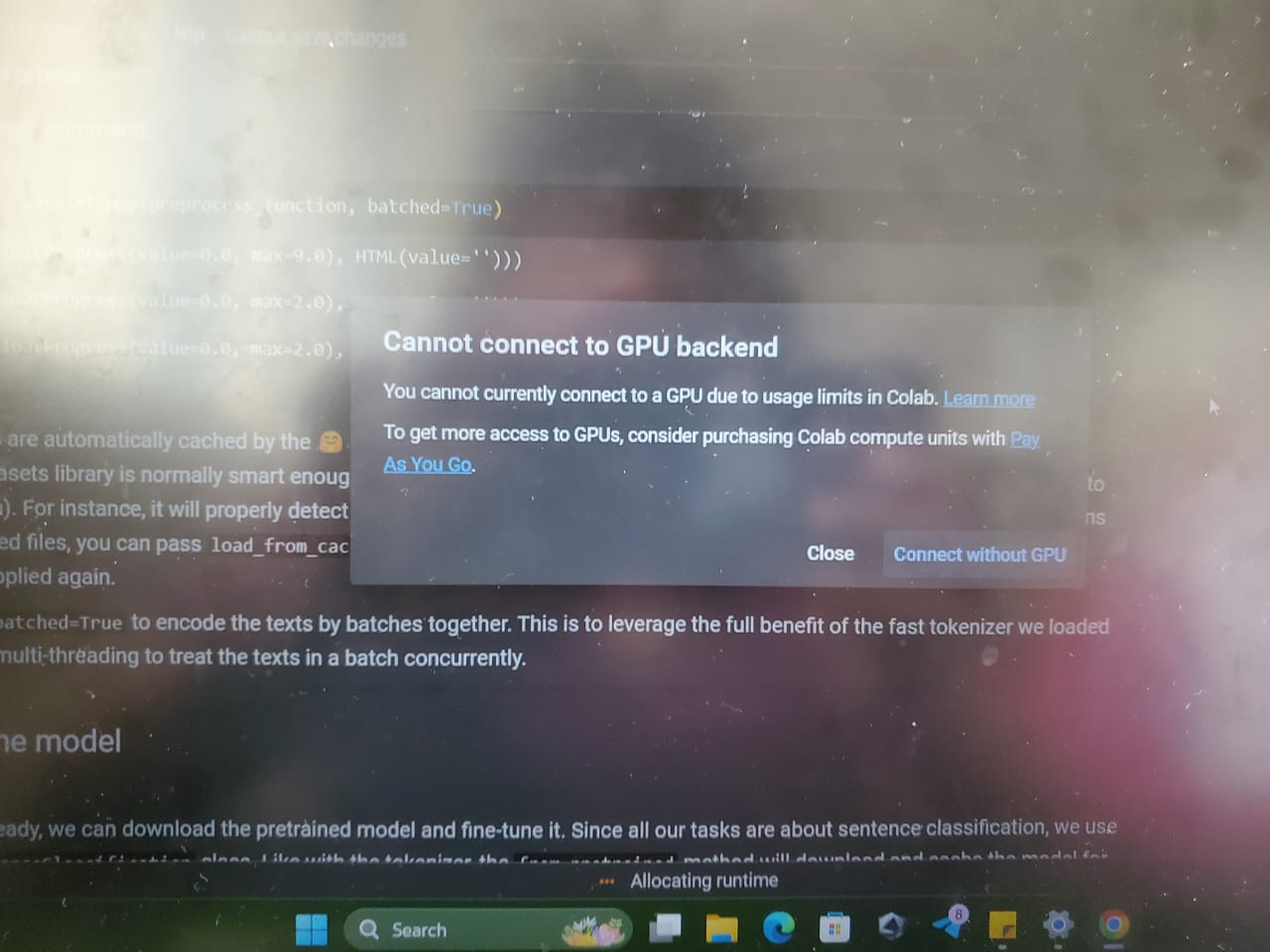
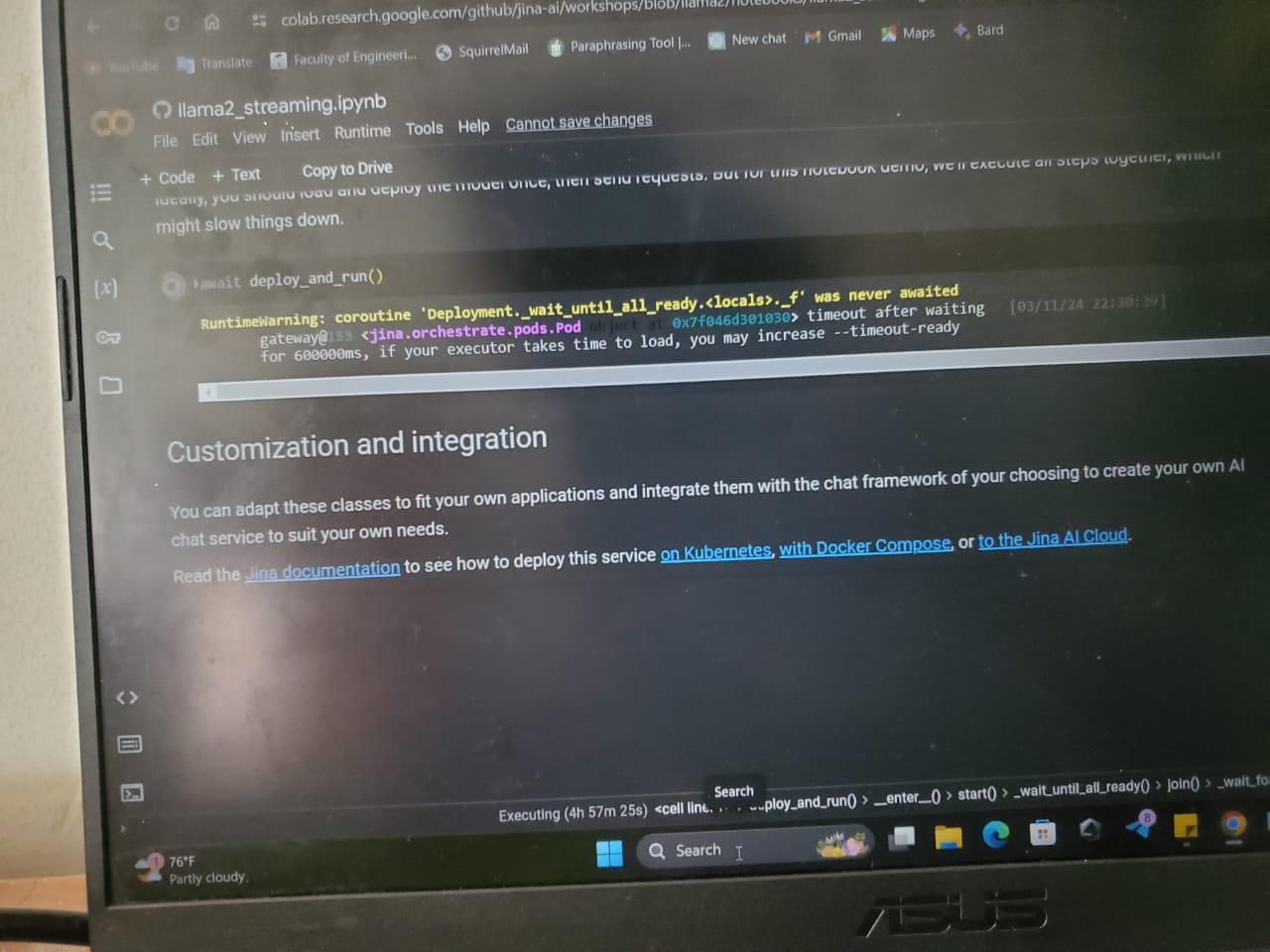
After we got access to some of the models of Llama 2

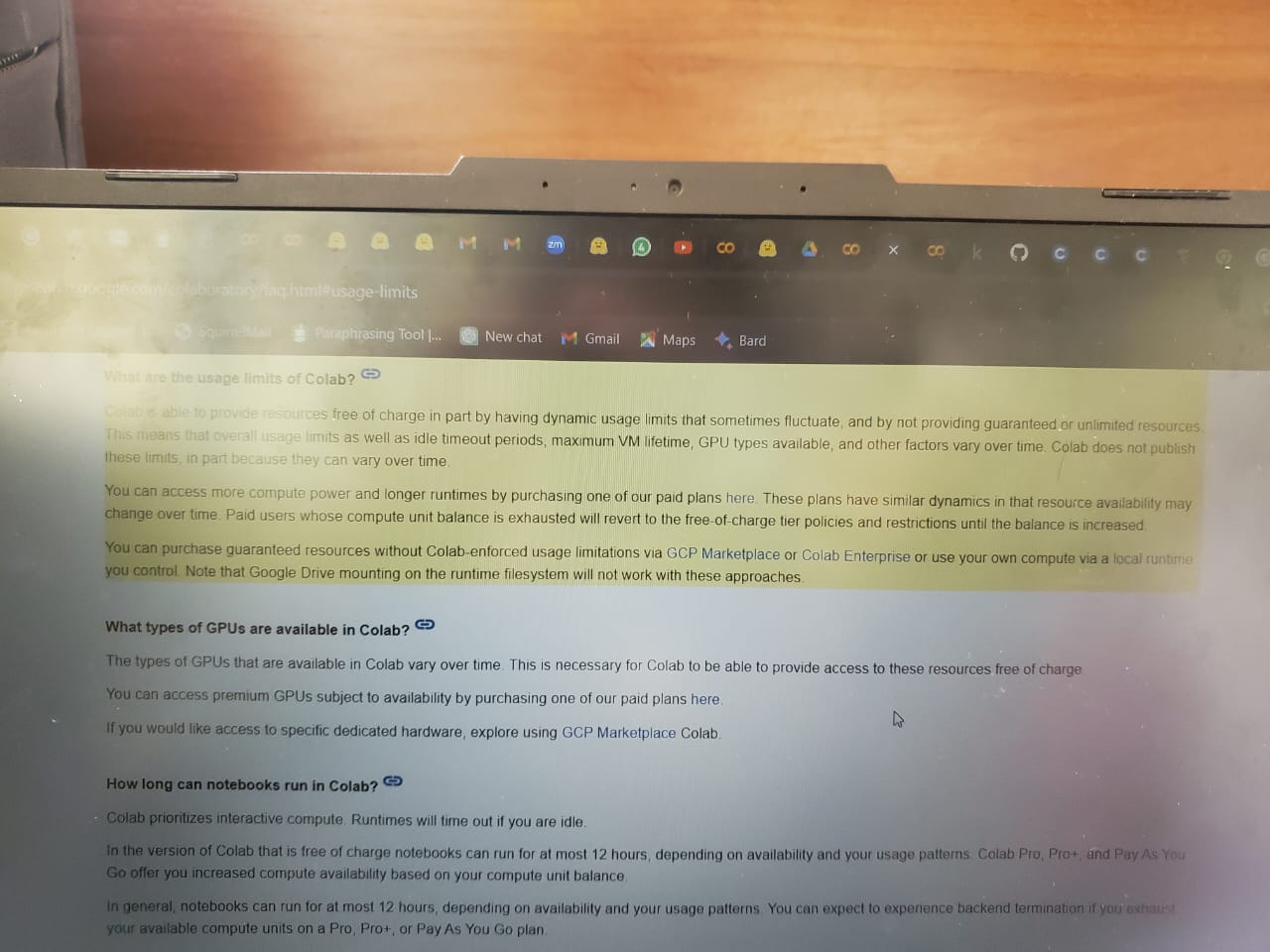
We selected the model Llama-2-7b and installed it on google colab with Huggingface.

Google Colab is only allowed with 15GB free space usage limit with GPU. But our original version of the model Llama-2-7b is 13GB, so we were not able to work with the model later. Then it automatically disconnected the GPU and started to work on the CPU.



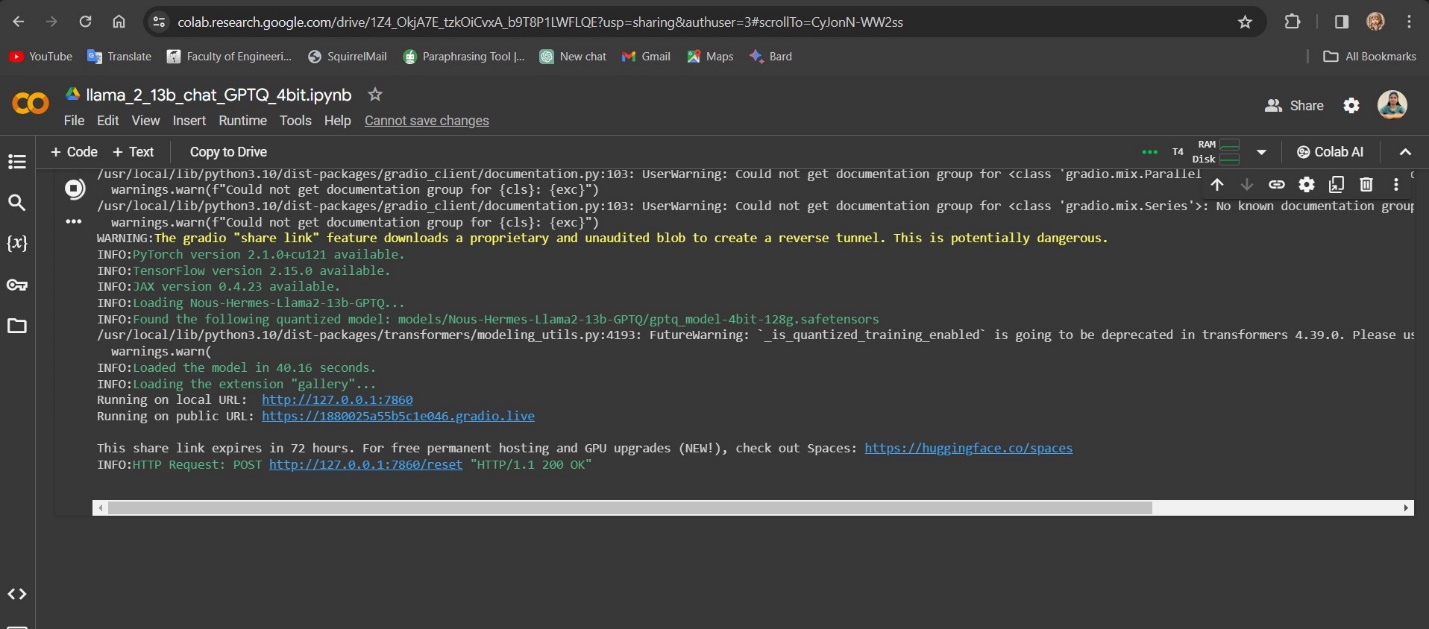
The capacity of CUP is not enough to run the model. When it was running on CPU it took more than 5 hours to launch the model, even it could not finish launching. You can find the time which it was taken while it was launching in the down part of this figure.

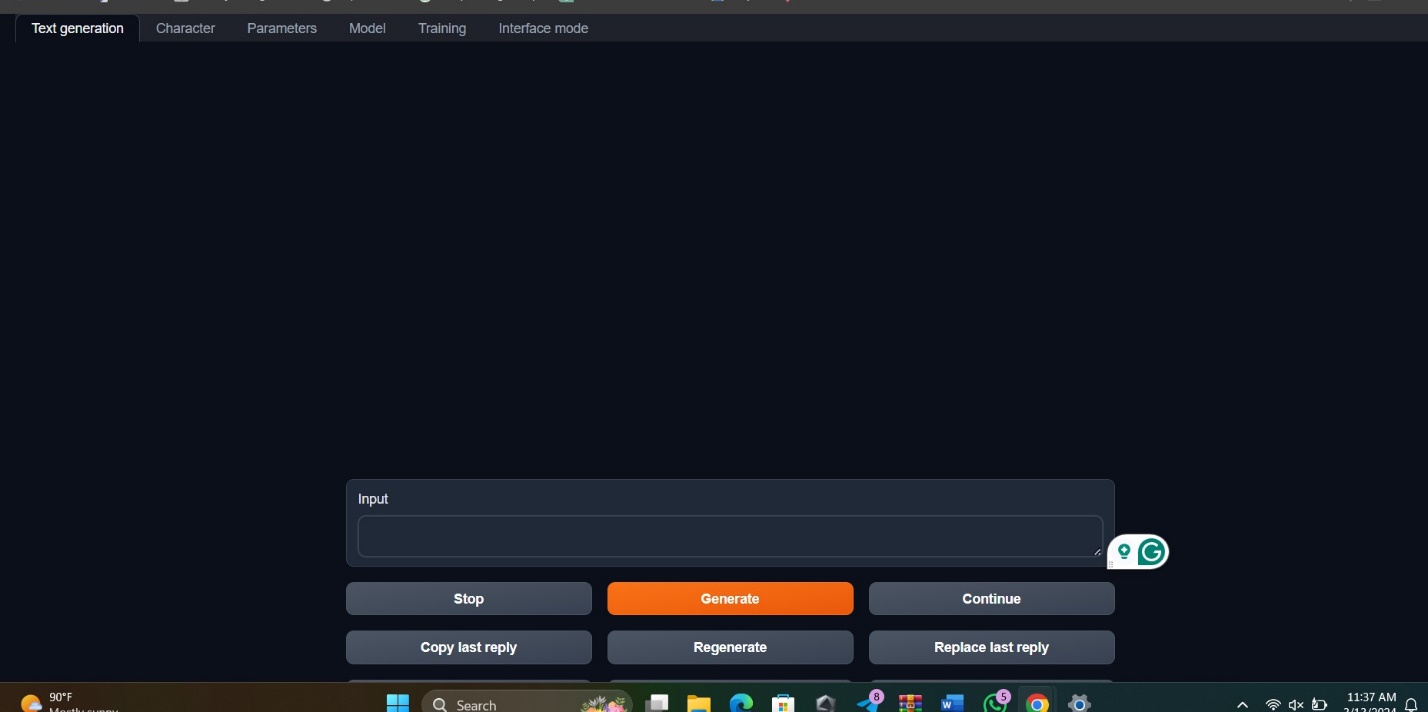


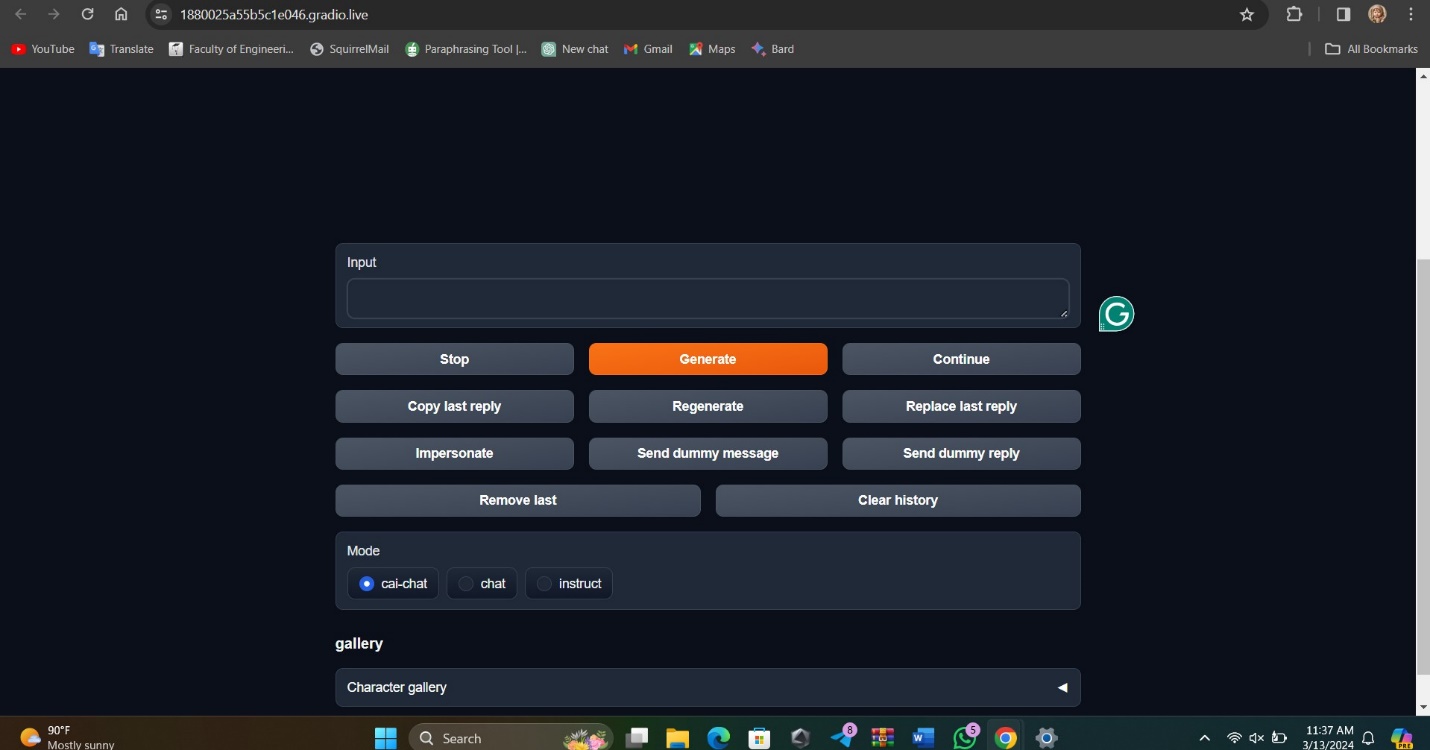
Then only we got to know about the limitations of free Google Colab.

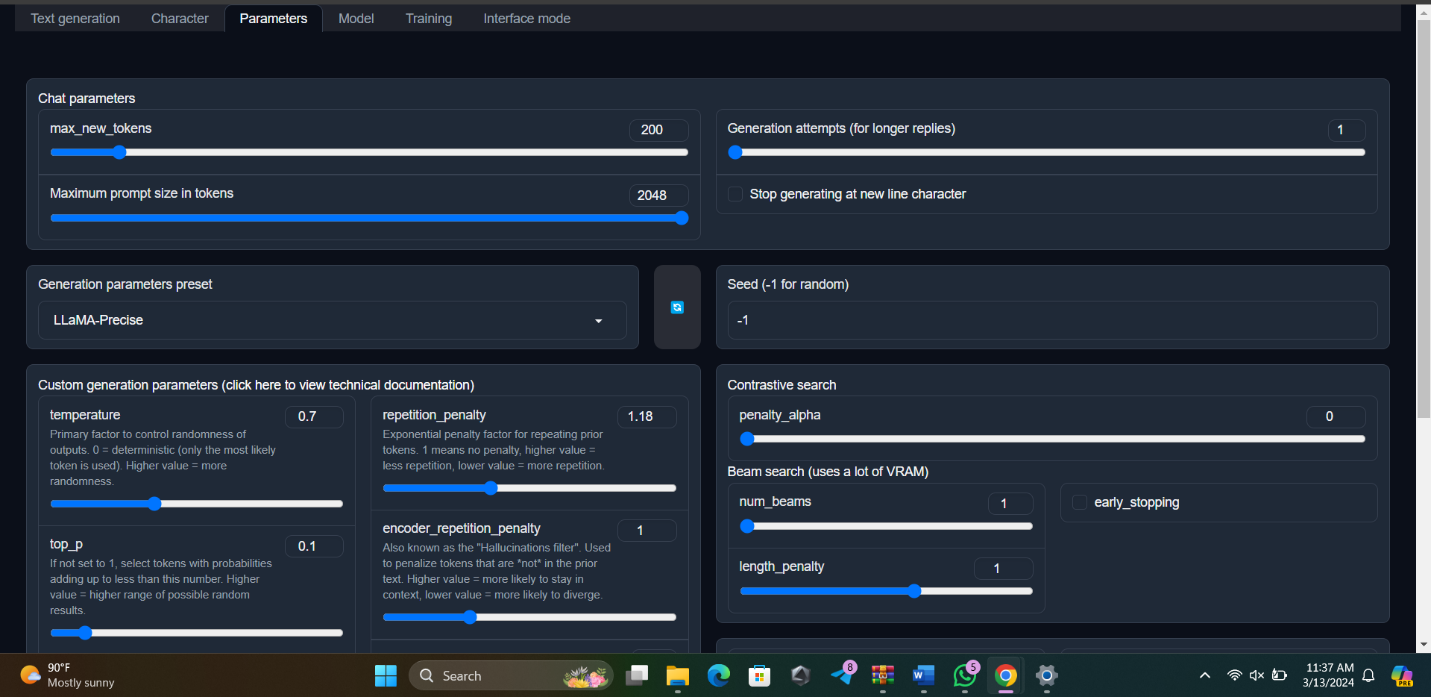
Then we have tried the Quantized version of Lllama-2 with only 5GB,

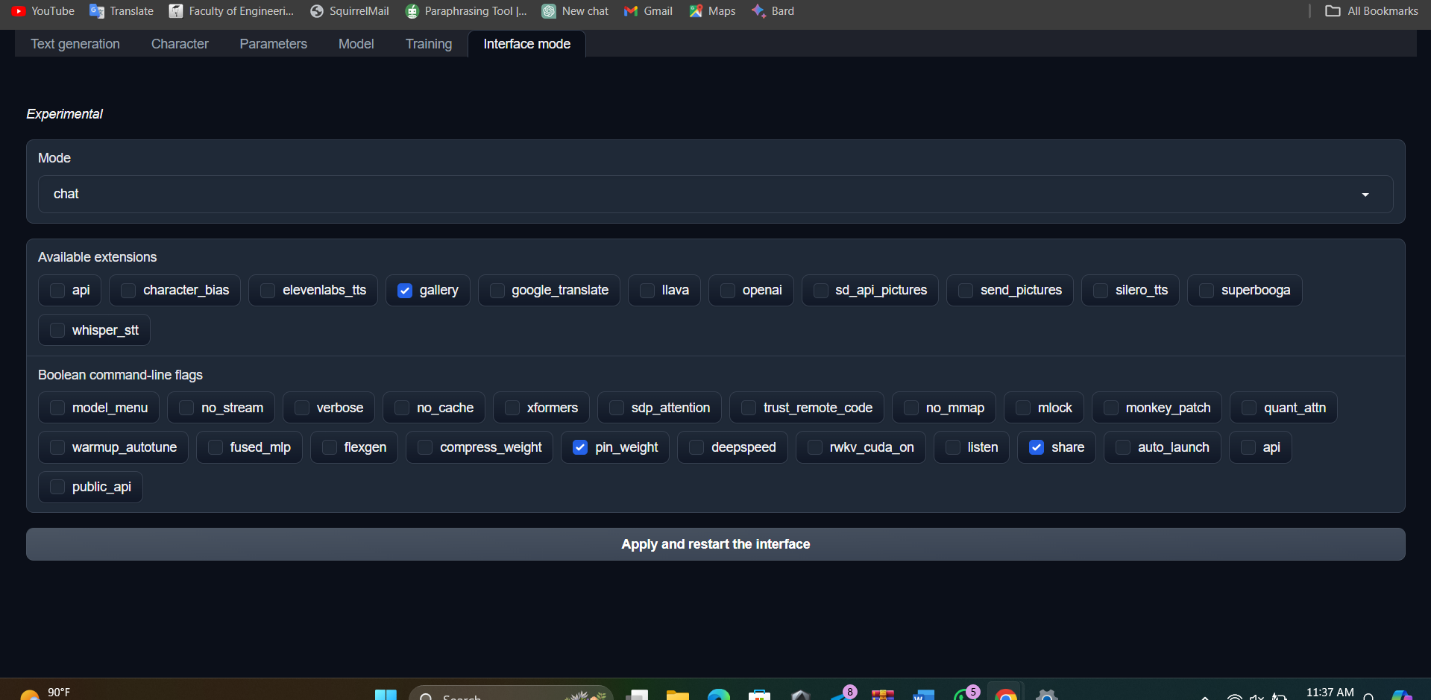
Then we could able to run the model on Google colab

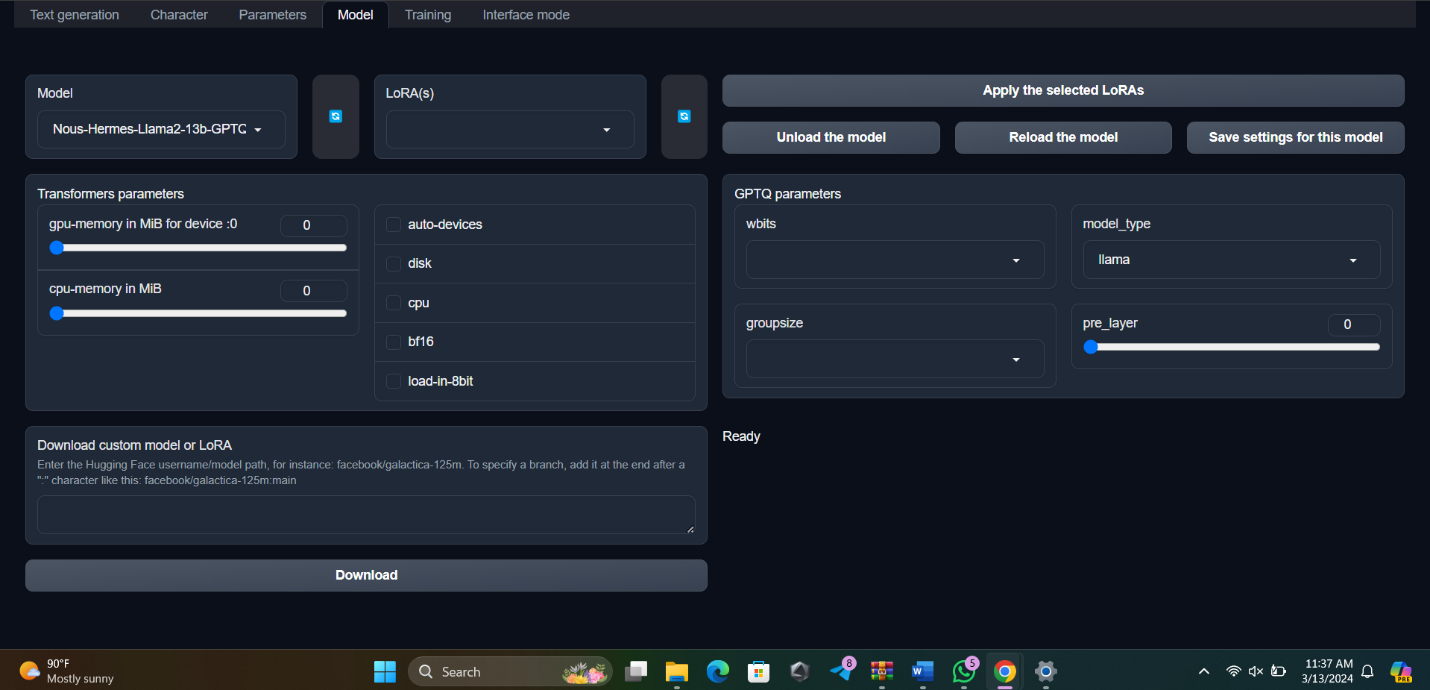


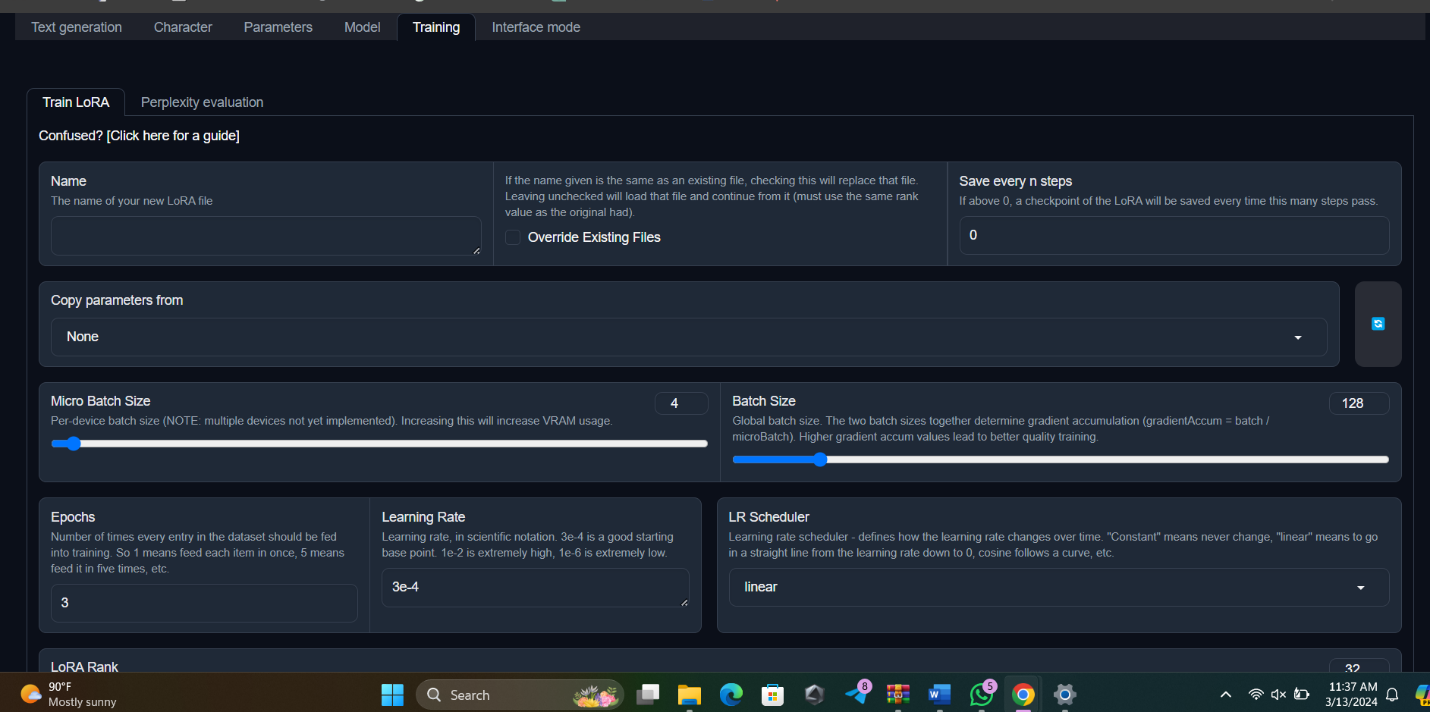












Now learning and working with LoRa fine-tuning.

Meanwhile, we have tried to launch the BERT model, that was sent by Farina. For that also our free colab allocation is not enough

