LLAMA -3 8B and 70B

The quantization process for each of the Llama 3 8B and 70B models has been tested separately on Google Colab and Kaggle. All methods were tried when converting the models to 8bit. These methods are AutoAWQ (Automatic Weight Quantization), Post-training Quantization, Quantization Aware Training, Dynamic Quantization, Tensor Cores - TensorRT, and BitsAndBytes. Of these methods, BitsAndBytes was the only one that worked. Details about this are in the file.

On the other hand, Llama 3- 8B 16 bits, Llama 3- 70B 8 bits and Llama 3- 70B 16 bits were unsuccessful. Finally, it was found that the model had a slow response time even when asked a simple question. I think the most reliable method is Llama 3- 8B 4 bit.

A screenshot of a computer

Description automatically generated

