## Assignment-1

Load Zephyr using GPTQ method and make some inferencing

GPTQ is a post-training quantization (PTQ) method for 4-bit quantization that focuses primarily on \*\*GPU\*\* inference and performance.

The idea behind the method is that it will try to compress all weights to a 4-bit quantization by minimizing the mean squared error to that weight. During inference, it will dynamically dequantize its weights to float16 for improved performance whilst keeping memory low.

## Assignment-2

Load Zephyr using GGUF method and make some inferencing