**ASSIGNMENT LLMs**

**Part-1**

**Answer the following questions in a google colab or jupyter notebook environment:**

1. Explain Context window of LLMs in a few lines and why this metric is important
2. Explain the self attention mechanism in LLMs
3. Consider you have an enterprise data you need any LLM to answer from, what are all techniques available to achieve this?
4. Explain what is meant by Quantization of a LLM
5. After an LLM has been pre trained, which procedure it goes through to Give Answer in a Q nd A format?

**Part-2**

**Your task is to use STT Whisper model to convert an audio file to Text: (Use Hugging face Transformers Library)**

1. Install Python and Required Libraries:
   * Install the necessary libraries: transformers, torch, and soundfile.
2. Load the Whisper Small Model:
   * Use the Hugging Face Transformers library to load the Whisper Small model and processor.
3. Load and Preprocess the Audio File:
   * Read your audio file (preferably in .wav format) and preprocess it using the processor.
4. Generate Transcriptions:
   * Use the loaded model to generate the transcription from the preprocessed audio.
5. Output the Transcription:
   * Decode and print the generated transcription.