Voice assistant for cell-phone tower technicians

Aim: To create a text to text based module for processing and answering user queries related to technician assistance.

This module would comprise of a LLM based solution, fine-tuned on user specific data for question answering capability

Building the solution can be divided into the following steps:

1.) Identify the type of user data
2.) Prepare Data for pre-training of the language model
3.) Build the model architecture based on the existing state of the art LLM architectures such as GPT, LLAMA
4.) Identify the model size- this is a very important step to make sure there are no latency issues during user operations and depends on the GPU compute constraints
5.) Pre-training of the model on the initial user data to make it capable of generating text
6.) Fine-tuning- this step involves making the model specific to our use case of giving it the question answering capability by Instruction-fine tuning
7.) Once the basic prototype of this module is ready, it can modified for handling out of the box data and learning ability by implementing RAG and RLHF.

Specifics on Model architecture:

1.) The model architecture has been adapted from the GPT model. It is a decoder only architecture meant for causal attention and causal text generation
2.) The model size according to the recent standards of LLM size is small (roughly 124M parameters) while large model parameters run into Billions!
3.) The model size has been set taking into consideration the inference speed and latency of the overall application also including text to speech and speech to text modules.
4.) In case of GPU constraints: We are thinking of using pre-trained weights for our architecture and can instead proceed with fine tuning our model using the current compute. Fine-tuning requires much lesser compute compared to pre-training.

Specifics on the Data preprocessing:

1.) For pre-training, data doesn't have to be supervised. The text is first broken down into small chunks by tokenization.

2.) To make sure our method is up to date with the current tokenization methods, we use BPE (Byte Pair Encoding) for our module. (GPT and LLAMA use the same).
3.) The main task is to prepare the fine-tuning data for instruction based fine tuning. A series of question and their respective answers must be prepared for the same.

Work done till now:

1.) I have created the data-loaders and preprocessing scripts for pre-training data
2.) The model architecture scripts have been written

Next steps to be taken:

1.) Training script has to be written both for pre-training and fine-tuning
2.) After all the scripts are written the training process can be commenced
3.) Test the initial model with sample user queries

All the work has been done in python and py-torch framwework has been used for building the model architecture.

We propose to use Lang-chain framework for modifying the basic prototype to implement RAG and RLHF for making it deployment ready.