

README

Chaabi Assignment

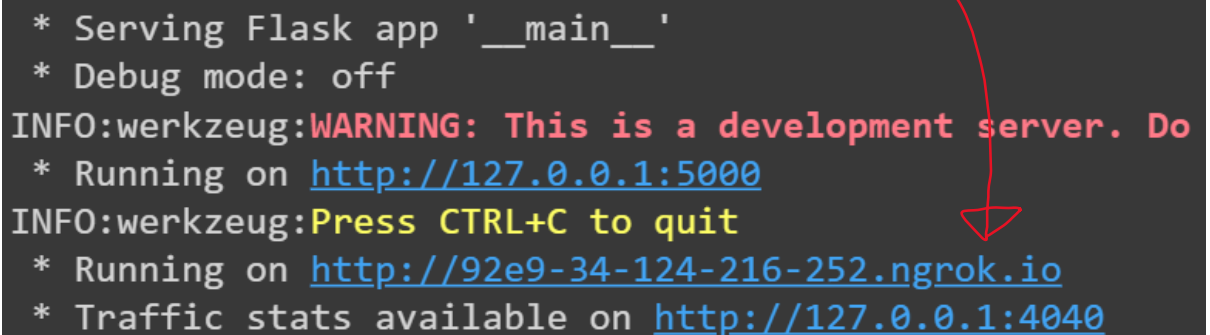
Made by:
Bhavneek Singh
2020EE10482
IIT Delhi

Introduction

For the assignment, I am using an Open Source LLM called “Mistral” it has 7B parameters, is fully open source, and performs at-par with Llama-2 13B (model used is 4-bit quantized in ‘gguf’ format for CPU inference). VectorDB used in the project is ChromaDB with Langchain

Instructions

1. Import the .ipynb notebook into Google Colab (T4 GPU Runtime)
2. From the ‘Runtime’ dropdown menu, select ‘Run All.’
3. Wait for a few minutes till the last cell is running properly
4. Copy the link just above the “Traffic stats...” in front of “Running on “



```
* Serving Flask app '__main__'
* Debug mode: off
INFO:werkzeug:WARNING: This is a development server. Do
* Running on http://127.0.0.1:5000
INFO:werkzeug:Press CTRL+C to quit
* Running on http://92e9-34-124-216-252.ngrok.io
* Traffic stats available on http://127.0.0.1:4040
```

5. Paste the link in the browser, put a ‘/’ at the end if not already there, then a ‘?question=’ and your query with ‘_’ as spaces, press enter, and wait for a minute for the results, here are some samples:

