Date of Submission: 12/07/2024

Name: Nipun Gupta

Institute: SRM Institute of Science and Technology Kattankulathur

Branch/Specialization: M.Tech(Integrated)- Artificial Intelligence [Dept: Computational Intelligence]

Registration Number: RA2312701010025

Internal Mentor: Dr.Sumathy G.

External Mentor: Dr. Vasudha Kumari (AI Software Solutions Engineer, Intel)

Introduction to GenAI and Simple LLM Inference on CPU and finetuning of LLM Model to create a Custom Chatbot

Project Overview

This project aimed to explore the concepts of Generative AI (GenAI) and demonstrate how to perform simple inference using a pre-trained GPT-2 model on a CPU. Additionally, the project showcased the process of finetuning a large language model (LLM) to create a custom chatbot tailored to specific conversational data.

Key Objectives

1. Understand the basics of GenAI and LLMs
2. Perform simple inference using a pre-trained GPT-2 model on a CPU
3. Finetune a GPT-2 model on a dataset to create a custom chatbot

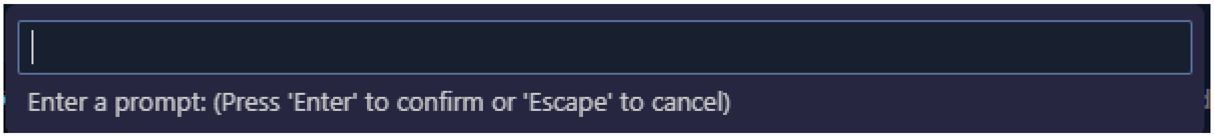
Methodology

1. Loaded the pre-trained GPT-2 model and tokenizer using the Hugging Face Transformers library
2. Set the device to CPU for inference
3. Defined a function to generate text based on a given prompt
4. Loaded a dataset for finetuning the chatbot model
5. Preprocessed the dataset by tokenizing the input text and creating target labels
6. Defined the training arguments and created a Trainer object
7. Fine-tuned the GPT-2 model on the dataset
8. Saved the finetuned chatbot model for future use

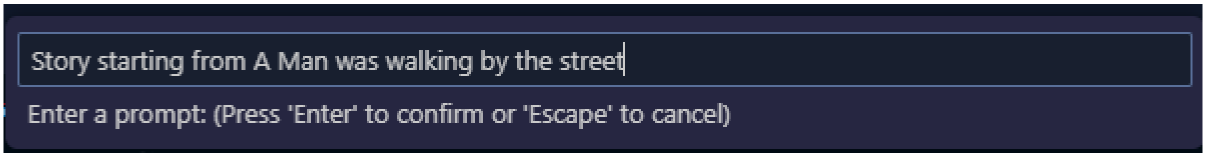
Results

1. **Successfully performed simple inference using the pre-trained GPT-2 model on a CPU**
2. **Generated coherent and relevant text based on the provided prompts**
3. **Finetuned the GPT-2 model on a dataset of conversational data**
4. **Saved the finetuned chatbot model for future use in generating responses for the custom chatbot**

Output

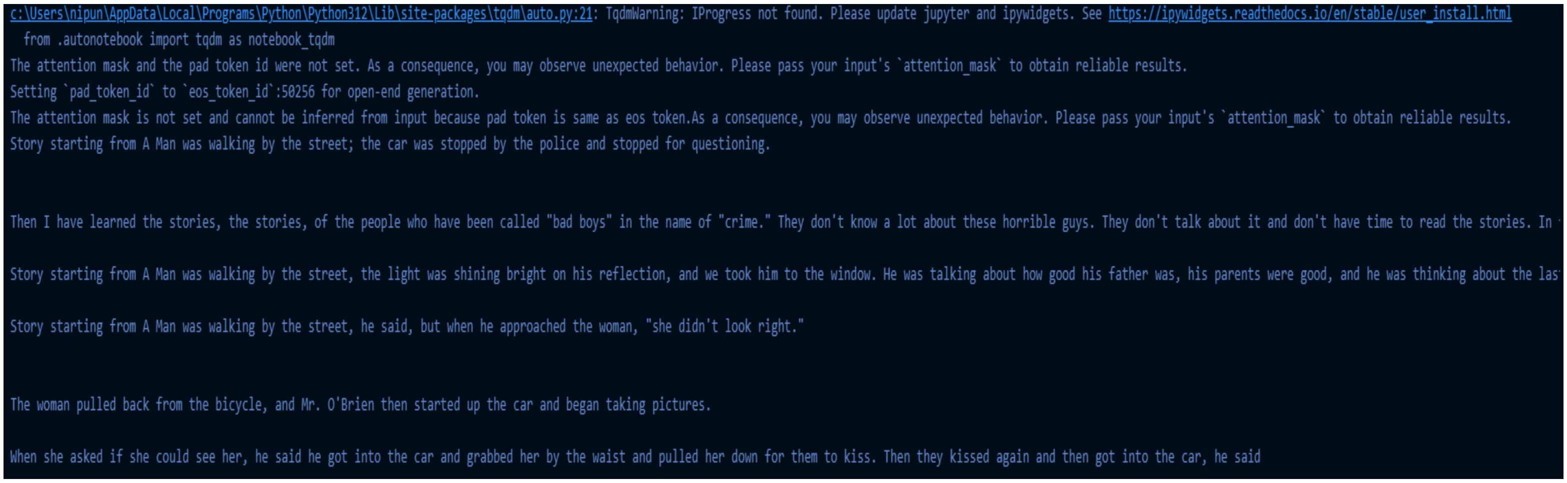


Enter a prompt: (Press 'Enter' to confirm or 'Escape' to cancel)



Story starting from A Man was walking by the street

Enter a prompt: (Press 'Enter' to confirm or 'Escape' to cancel)



```
c:\Users\nipun\AppData\Local\Programs\Python\Python312\Lib\site-packages\tqdm\auto.py:21: TqdmWarning: IProgress not found. Please update jupyter and ipywidgets. See https://ipywidgets.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask` to obtain reliable results.
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
The attention mask is not set and cannot be inferred from input because pad token is same as eos token.As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask` to obtain reliable results.
Story starting from A Man was walking by the street; the car was stopped by the police and stopped for questioning.


Then I have learned the stories, the stories, of the people who have been called "bad boys" in the name of "crime." They don't know a lot about these horrible guys. They don't talk about it and don't have time to read the stories. In

Story starting from A Man was walking by the street, the light was shining bright on his reflection, and we took him to the window. He was talking about how good his father was, his parents were good, and he was thinking about the las

Story starting from A Man was walking by the street, he said, but when he approached the woman, "she didn't look right."


The woman pulled back from the bicycle, and Mr. O'Brien then started up the car and began taking pictures.

When she asked if she could see her, he said he got into the car and grabbed her by the waist and pulled her down for them to kiss. Then they kissed again and then got into the car, he said
```

Conclusion

This project provided an introduction to GenAI and demonstrated the process of performing inference using a pre-trained LLM on a CPU. It also showcased the finetuning of an LLM to create a custom chatbot tailored to specific conversational data. The finetuned model can now be used to generate relevant and coherent responses for the chatbot, making it a valuable tool for various applications.

Future Improvements

1. **Explore the use of different LLMs, such as GPT-3, for inference and finetuning**
2. **Optimize the finetuning process by experimenting with different hyperparameters and training strategies**
3. **Integrate the finetuned chatbot model into a larger application or system**
4. **Evaluate the performance of the chatbot using relevant metrics and user feedback**