High Radius Gen AI Assignment

by Anshul Chandra

Fine Tuning

- Model used: `llama-2-7b-chat-hf`
- The dataset needs to be transformed in the right format. It is currently in a .csv file consisting of a table with 2 rows corresponding to a Human Prompt and a Model Answer. The correct format is as follows: `<s>[INST] <<SYS>> System Prompt <</SYS>> User Prompt [/INST] Model Answer </s>`
- Link to transformed data: https://huggingface.co/datasets/asucada/ghl-support-docs
- Free Google Colab offers a 15GB Graphics Card (Limited Resources → Barely enough to store Llama 2–7b's weights)
- We also need to consider the overhead due to optimizer states, gradients, and forward activations
- To drastically reduce the VRAM usage, we must fine-tune the model in 4-bit precision, which is why we'll use QLoRA here.
- QLoRA with a rank of 64 and a scaling parameter of 16 is being leveraged. We'll load the Llama 2 model directly in 4-bit precision using the NF4 type and train it for 2 epochs
- Link to fine-tuned model and tokenizer on HF: https://huggingface.co/asucada/Llama-2-7b-chat-finetune/tree/main
- How to improve the fine-tuning process?
    - Increase the size of the dataset
    - Work with a larger model such as `Llama-2-70b-hf`
    - Play around with hyperparameters like epoch, learning rate
    - Use a validation set while training

RAG Implementation

- Fine-tuned model has been used: `Llama-2-7b-chat-finetune`
- Embedding model: `all-mpnet-base-v2`
- Retriever: `Chromadb`
- Chunker: `Semantic Chunking`
- System Prompt: this takes in the context provided along with the query and gives the LLM a framework to follow and generate a response

```
"""[INST] <<SYS>>
You are a trained support bot to guide people about a SaaS pr
oduct. You will answer user's query with your knowledge and t
he context provided.
If a question does not make any sense, or is not factually co
herent, explain why instead of answering something not correc
t. If you don't know the answer to a question, please don't s
hare false information.
Do not say thank you and tell you are an AI Assistant and be
open about everything.
<</SYS>>
Use the following pieces of context to answer the users quest
ion.
Context : {context}
Question : {question}
Answer : [/INST]
"""
```

- Only the `text` column of the dataset has been used after being converted to PDF format

RAG Evaluation

Ragas is used to evaluate the efficiency of the generated outputs. Some of the metrics being tracked include:

- Retrieval metrics
  - Context Precision: compares question and context
  - Context Recall: compares ground truth and context
- Generation metrics
  - Answer Relevance: compares question and answer
  - Faithfulness: compares answer and context

I have manually prepared the following queries & ground truth for evaluating models by using ChatGPT.

```
queries = [
    "How to import a Funnel From ClickFunnels?",
    "How To Redirect HighLevel Domains?",
    "How To Build Courses / Membership Sites?",
    "How to create invoices in HighLevel?",
    "How to Setup SFTP for your WordPress Site?",
    "How to configure Staff Member selection in Round Robin Cale
    "How to link privacy policy when you collect personal data f
    "How to Use Webhooks in HighLevel (Zapier)?",
    "Where is my affiliate payout?",
    "How to track leads from an Affiliate Campaign (Forms, Surve

]

ground_truths = [
    "To import a funnel from ClickFunnels to GoHighLevel, access
    "To redirect HighLevel domains with path, transfer the domai
    "To build HighLevel Membership/Course websites, access the c
    "To create invoices in HighLevel, navigate to the Payments t
    "To securely access the files and folders of your WordPress
    "To configure Staff Member selection in Round Robin Calendar
    "To link your privacy policy when collecting personal data,
    "To use webhooks in HighLevel, create a dedicated webhook UR
```

```
    "If you haven't received your affiliate payout, ensure you'
    "To track leads from Forms, Surveys, and Calendars in your a
]
```

Scores

	Base Model	Fine Tuned Model
Faithfulness	0.955000	
Answer Relevance	0.923192	
Context Precision	0.733333	
Context Recall	0.916667	
Harmfulness	0.000000	

Application

Note that the app is throwing the following error even-though the packages have been installed:

```
ImportError: Using bitsandbytes 8-bit quantization requires Acce
```

After some googling, I found out that it is because `bitsandbytes` is not completely support by Apple M1 chip. Regardless, if you own a Windows/Linux OS, please try to run the following:

The project is divided into 3 different Python files

1. `app.py` : Used to run streamlit application
2. `utils.py` : Create prompt template and retrieval QA chain
3. `ingest.py` : Create vector store and split document into chunks

Installation

1. Clone the repository

```
https://github.com/asucada/highradius.git
```

2. Create a virtual environment

```
python3 -m venv venv
```

3. Activate the virtual environment

```
source venv/bin/activate
```

4. Install the requirements

```
pip install -r requirements.txt
```

Usage

1. Ingesting Knowledge into the Chatbot.
    - Add your pdfs or docx files to the `dataset` folder.
    - Run the following command to ingest the knowledge into the chatbot.

    ```
    python ingest.py
    ```

    - This will create a `/vectorstore` folder which will contain the vectorized knowledge.
2. Running the Chatbot.
    - Run the following command to start the chatbot.

    ```
    python app.py
    ```