Spring 2025 | DATA 266 | Homework -3
Deadline – 11.59 PM – 04/23/2025
20 Points

Question # 1 (10 Points)

You are tasked with building an intelligent restaurant recommendation system that uses Retrieval-Augmented Generation (RAG) to answer free-form user queries. The underlying dataset includes detailed restaurant metadata such as description, review tags, features, location, hours, and cuisines.

```
import pandas as pd

df = pd.read_json("hf://datasets/itinerai/restaurants/restaurant_data.json")
```

You will build two versions of this system:

1. **Naive RAG**: Retrieval purely based on vector similarity.
2. **Re-ranking RAG**: Retrieval followed by a re-ranking step using an LLM or a cross-encoder.

System Architecture

1. **Document Creation**
   - Combine relevant fields into a single natural paragraph per restaurant.
   - Example:

     "The Consulate UWS is a French-American restaurant with brunch and vegetarian options. Located on the Upper West Side, it offers outdoor seating and receives excellent ratings for its atmosphere and pasta dishes."
2. **Naive RAG Pipeline**
   - Encode each document using a dense retriever
   - Use FAISS or similar to build an index.
   - Given a user query, retrieve top-3 restaurants and pass them + query to a generator
   - Generate a recommendation.
3. **Re-ranking RAG Pipeline**
   - Retrieve top-10 restaurants using the same retriever.
   - Re-rank them
   - Select top-3 re-ranked items and generate response.

Queries for Evaluation

Use the following 10 sample queries:

1. I want a cozy Italian restaurant near downtown that's open late.
2. Any vegan-friendly brunch spots with wheelchair accessible, outdoor seating and good reviews?
3. Where can I eat seafood near the beach with a romantic vibe?
4. Suggest a casual lunch place in the general area of Chinatown.
5. Are there any highly-rated spots for dessert after 10 PM?
6. I'm looking for a halal or vegetarian dinner place near the Riverwalk with highchair avaiable.
7. Recommend a mid-range price Mexican dog friendly restaurant with quick service.
8. Any hidden gems for coffee or breakfast with quiet seating?
9. Find a sushi place with scenic views, serves alcohol, and not too expensive.
10. I want to go somewhere trendy and lively for drinks and appetizers.

Evaluation: For each query, compare the Naive RAG and Re-ranking RAG outputs using the following criteria:

Criteria	Scoring Method
Relevance	Manual score from 1 to 5
Diversity	Do responses vary across queries?
Fluency	Language quality of the generated output

Question # 2 (10 Points)

Teaching One Model to Follow Instructions: Fine-Tuning with Dolly 15K

In this assignment, you will instruction-tune one decoder-only language model (LLM) using the **Dolly 15K** dataset, simulating how industry systems like ChatGPT are fine-tuned to follow user instructions. You will also evaluate the model's improvements in terms of helpfulness, fluency, and instruction adherence.

The fine-tuning must ensure that **only the response contributes to the loss function**, using response-only masking. You will use the Databricks Dolly 15K dataset. Load it using:

```
import pandas as pd
df = pd.read_json("hf://datasets/databricks/databricks-dolly-15k/databricks-dolly-15k.jsonl", lines=True)
```

Prompt Formatting

```
### Instruction:
<instruction>

### Context:
<context>  # Include only if not empty

### Response:
<response>
```

Requirements

- Use only one decoder-only model, such as:
  - gpt2
  - EleutherAI/gpt-neo-125M
  - tiiuae/falcon-rw-1b
  - mistralai/Mistral-7B (if your environment supports it)
- Use Hugging Face's transformers library for training.
- Use DataCollatorForCompletionOnlyLM to apply instruction masking**,** so that only the response tokens are used to compute loss**.**
- Train for 3-5 epochs on a filtered subset

Evaluation: Select 10 held-out instruction samples and compare the outputs from:

1. The **base (pretrained)** model

2. The **instruction-tuned** model

Use the following evaluation rubric (1–5 scale):

Metric	Description
Instruction Following	How well does the output follow the instruction?
Helpfulness	Is the response useful and relevant?
Fluency	Is the output grammatically and stylistically sound?

For **3 cases**, include a short reflection explaining what improved and why. Also include **1 failure case** and hypothesize the cause.

Bonus (Optional) **(4 Points)**

- Apply LoRA or QLoRA to train and compare the results with the fine-tuned model in terms of the above-mentioned evaluation.

You are required to follow:

1. Submit **one** MS/PDF/Scanned document:
   - Include all the steps of your calculations.
   - Include the summary of the model.
   - Attach screenshots of your code.
   - Attach screenshots – showing first few epochs of model training.
   - Attach screenshots of the important code outputs such as confusion matrices, learning curves, and classification reports.
2. Source code:
   a. Python (Jupyter Notebook)
   b. Ensure it is well-organized with comments and proper indentation.

- **Failure to submit the source code will result in a deduction of full/partial points.**
- Format your filenames as follows: "your_last_name_HW1.pdf" for the document and "your_last_name_HW1_source_code.ipynb" for the source code.
- Before submitting the source code, please double-check that it runs without any errors.
- Must submit the files separately.
- Do not compress into a zip file.
- HW submitted more than 24 hours late will not be accepted for credit.