CS 5393

Midterm Exam
Due: Thursday Apr. 10

Exploring Open-Source LLMs with Ollama

## Objective

To gain hands-on experience with open-source Large Language Models (LLMs) using Ollama, and to develop a deeper understanding of their capabilities, limitations, and potential applications.

## Background

Ollama is a tool that simplifies running open-source LLMs locally. It provides an easy way to download, run, and experiment with various models from Hugging Face and other sources.

## Tasks

### 1. Setup and Model Selection

a) Install Ollama on your local machine or SMU genuse server.

b) Choose three diferent open-source LLMs available through Ollama. Select models with varying sizes and architectures (e.g., a small model like TinyLlama, a medium-sized model like Mistral-7B, and a larger model like Llama2-70B).

c) Download and set up these models using Ollama.

### 2. Basic Model Exploration

For each of the three chosen models:

a) Perform a series of basic tasks:

- **General question answering**
- **Text summarization**
- **Simple code generation**
- **Creative writing (e.g., a short story prompt)**

b) Document the prompts used and the outputs generated.

c) Compare the performance of the models in terms of:

- **Response quality**
- **Speed of generation**
- **Resource usage on your machine**

### 3. Focused Experimentation

Choose one area of focus from the following:

a) Prompt Engineering:

- Experiment with different prompting techniques (e.g., few-shot, chain-of-thought, self-consistency) across your chosen models.
- Analyze how these techniques affect the output quality and consistency.

b) Domain-Specific Performance:

- Select a specific domain (e.g., medical, legal, technical documentation).
- Create a set of domain-specific tasks and evaluate how each model performs.

c) Multilingual Capabilities:

- Test the models' performance across at least three different languages.
- Evaluate their translation abilities and understanding of language-specific nuances.

d) Ethical Considerations:

- Design a series of prompts to test for biases or potentially harmful outputs.
- Analyze how different models handle ethically sensitive queries.

### 4. Analysis and Report

Write a comprehensive report (2000-2500 words) that includes:

a) An overview of the models you chose and why.

b) Detailed results from your basic exploration and focused experimentation.

c) Analysis of the strengths and weaknesses of each model.

d) Insights gained about open-source LLMs and their capabilities.

e) Reflections on the practical implications of your findings for real-world applications.

f) Discussion of any challenges faced during the assignment and how you overcame them.

## Deliverables

1. A GitHub repository containing:

   - All code used for running models and experiments

2. The comprehensive report in PDF format

3. A 10-minute presentation summarizing your findings (to be delivered in class)

## Evaluation Criteria

- Thoroughness of model exploration and experimentation (35%)

- Quality and depth of analysis in the report (25%)

- Creativity in experiment design and problem-solving (15%)

- Clarity and organization of code, documentation, and presentation (25%)

## Resources

- Ollama Oficial Documentation: https://ollama.ai/docs

- Hugging Face Model Hub: https://huggingface.co/models