Course: EECS4080 | **Term**: Summer 2025
Project title: Evaluating Large Language Models on Code Behavior and Execution Analysis

Supervisor name (email): Song Wang (wangsong@yorku.ca)
Student name: Somaiyah Sultani |**Student number**: 213967534
Student's email: ssultani@my.yorku.ca

Project Description

This project aims to evaluate the capabilities of Large Language Models (LLMs) in understanding and analyzing code behavior based on execution results. While LLMs have shown strong performance in code generation and completion, their ability to reason about dynamic execution, such as interpreting outputs, diagnosing runtime errors, and explaining unexpected behaviors, in general, remains underexplored. We will develop a benchmark dataset containing code snippets paired with execution outcomes (e.g., outputs, errors, return values) and assess LLMs on tasks including output prediction, behavior explanation, and error diagnosis. The evaluation will consider both quantitative metrics (e.g., accuracy) and qualitative aspects (e.g., reasoning depth), offering insights into the strengths and limitations of current LLMs in execution-aware code analysis.

Course Learning Outcomes

1. Apply the knowledge they have gained in other computer science courses to a real-world system.
2. Articulate the questions that a particular area of research in computer science attempts to address.
3. Conduct independent research in some aspect of computer science.
4. Prepare a professional presentation that outlines the contributions they made to the project and the knowledge acquired during the project
5. Knowing how to run LLMs for a specific SE task

Background Requirements

Prerequisites: minimum EECS GPA is 4.5

Resources

Course Website: https://wiki.eecs.yorku.ca/course_archive/2019-20/Y/4080/course_descriptions

Python Package Repository:

1. Pypi: https://pypi.org
2. API for downloading Pypi projects: https://pypi.org/simple/
   https://pypi.org/pypi/p/json/

LLMs:

1. https://huggingface.co/LLMs
2. LLaMA Website: https://ai.facebook.com/blog/large-language-model-llama-meta-ai/
3. Alpaca Website: https://crfm.stanford.edu/2023/03/13/alpaca.html

Python IDE:

Pycharm

Experiment Data:

LiveCodeBench:

https://livecodebench.github.io/leaderboard.html
https://github.com/LiveCodeBench/LiveCodeBench
https://huggingface.co/livecodebench

Readings

1. An Empirical Evaluation of Using Large Language Models for Automated Unit Test Generation
   https://arxiv.org/pdf/2302.06527.pdf
2. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code
   https://arxiv.org/abs/2403.07974

Deliverables

The list of artifacts to be completed by the final due date; includes programs, documentation, reports, user guides, etc. One of them is your final project poster presentation on a date around the end of the term.

Breakdown

Week	Content Covered
1	**Background & Related work reading/writing**: Paper reading & understanding background knowledge

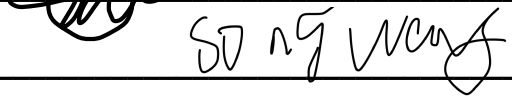
2	**Background & Related work reading/writing**: Paper reading & understanding background knowledge; working on the preliminary project description document
3	**Data Collection**: Collect essential data tools and information on given projects;
4	**Data Collection**: Collect essential data tools and information on given projects;
5	**Experiment Setting Up**: Set up LLM environment
6	**Experiment**: Use existing LLMs to generate data for the experimental projects.
7	**Experiment**: Use existing LLMs to generate data for the experimental projects.
8	**Result Collection**: Collection of the results of LLMs;
9	**Result Collection**: Collection of the results of LLMs;
10	**Result Analysis**: Compare the results and evaluate LLMs;
11	**Result Analysis**: Compare the results and evaluate LLMs
12	**Report writing:** Wrap up the project and finish the final presentation
13	**Report writing**: Wrap up the project and finish the final report
14	**Project Presentation**

Evaluation

- Mid-project status report – 15%
- Final documentation & reports – 25%
- Programming and technical work – 20%
- Poster/presentation – 30% (required component)
- Meeting milestones – 10%

Signatures

Student 

Supervisor 

Course Coordinator _____