[LLM Agents Course] Release of Labs 1, 2, 3

The lab assignments for this course have been released, and are due by **December 12, 11:59pm PST.** The first lab will require you to use the Autogen framework to build a pipeline for analyzing restaurant reviews, assigning a score to each restaurant. The second and third labs relate to prompt engineering for LLM security. In lab 2, you will be writing attack prompts while in lab 3, you will be writing defense prompts.

Here are google drive links to a `.zip` file for each lab. Please download the `.zip` file to obtain the starter code for each lab.

- Lab 01, Restaurant Review Pipeline. Link.
- Lab 02, Writing LLM Attacks. Link.
- Lab 03, Defending Against LLM Attacks. Link.

Asking Questions, Collaboration, and FAQ

We strongly recommend using Discord to communicate with course staff and ask any questions related to the lab. We will be monitoring Discord and will be actively responding to questions. In addition, we have this ongoing FAQ document. Before posting, please refer to the FAQ document to see if the question has already been addressed.

When collaborating with others, avoid giving your exact solution. We encourage conceptual discussion, but please avoid leaking any specific implementation details and/or code files.

Completion and Submission

For lab 01, we provide a set of public tests within the file `test.py`. Use these to verify your solution. Passing these will be an indicator that your solution is complete. Submit your completed lab implementations to the following Google Forms:

- Lab 1 Submission Google Form
- Lab 2 Submission Google Form
- Lab 3 Submission Google Form

Sometime after submission, we will be emailing you the result of whether or not you successfully completed the lab. The threshold for passing the lab is defined as the following:

- Lab01: Passing 3/4 of the public tests
- Lab02: Passing at least 1/2 of the hidden tests. The first hidden test is essentially the same as the mock system message provided in the image of the OpenAI Playground in `Instructions.md`.
- Lab03: Passing at least 1/3 of the hidden attack tests. If your defense works against your attacks written in Lab02, you should have no problem getting credit.

OpenAI API Keys

Please use your own OpenAI API Keys through the following portal: https://platform.openai.com/docs/quickstart. For all labs, please use the GPT-4o-mini model. We expect the anticipated cost to be < $1 per person.