Take Home Exercise

Your task is to create an internal dashboard for an evaluation server comparing two different models on a range of performance benchmarks.

Specifically, you will evaluate the **models** DeepSeek-Coder and CodeGemma on the following **benchmarks**:

- HumanEval
- MBPP
- Multipl-E

using the **Pass@k metric** for k=1, 3, 5.

For the dashboard, emphasize on functionality and ease of use. Eg, you can have buttons to toggle models, benchmarks, the parameter k, and any other necessary parameter. We often care about comparing different settings, so the ability to visualize multiple settings at the same time is a bonus. Finally, use your own judgement on best visualization techniques for different options [charts, buttons, etc].

Deliverable:

A Github repo with all the code and README containing instructions to:

- install the conda environment
- command to perform the evals for a given configuration of model, k, benchmark (offline)

- script to run a sweep for the evals for both models, all k's, all benchmarks (offline)
- display the results of the script on a dashboard ui (you can use any ui you like)
- a `tests` folder containing at least two tests that can be run via `pytest`

You can assume we will be running everything on an Ubuntu instance. If you have any questions, please reach out to us.

Notes:

- If you do not have access to a high end GPU, feel free to sub-select the first 50 examples for each benchmark. In case you do so, please mention that explicitly in the README.
- Some benchmarks like Multipl-E might have too many dependencies. Feel free to search for relevant Docker containers online if helpful.
- For the UI, feel free to use whatever tool you find most convenient (e.g., plotly or any other interface you are familiar with). The goal is to test quick hacking skills.
- Do not spend time over-optimizing fonts, color schemes, etc. As long as they are legible and well-defined (eg, legends), it is fine.
- Feel free to take help from code LLMs for any part of this assignment.