CS5787 Milestone Implementation Plan

For the preliminary milestone, our plan is to run a small-scale evaluation using existing data from LiveCodeBench. Since LCB already provides both model-generated code solutions and their corresponding unit-test results, we can directly sample around five to ten tasks and use the official unit-test outcomes as the objective correctness signal. For each selected task, we will take the candidate solutions produced by different LLMs, record their pass/fail results from the LCB unit tests, and treat these as the gold standard. We will then apply several LLMs as subjective judges by prompting them to evaluate the same candidate solutions without executing any code. By comparing the subjective judgments with the objective unit-test outcomes, we can compute simple agreement rates and observe how often LLMs correctly identify the better solution. This small experiment will give us initial insight into the reliability of LLM-as-a-Judge for coding tasks, help validate our evaluation pipeline, and provide early evidence on model tendencies or inconsistencies before we scale to larger datasets or introduce agent-based judges.