Home Assignments

Question 1

Please follow the steps below to serve the model Qwen/Qwen2.5-Coder-0.5B using vLLM or sglang (GPU or CPU) and evaluate its performance on the HumanEval(openai/human-eval: Code for the paper "Evaluating Large Language Models Trained on Code) task, check in all code/doc in a github repo and share with us:

- Serving the Model with vLLM or SgLang or **llama.cpp**:
  - Utilize the vLLM or SgLang or **llama.cpp** to serve the model on the CPU or GPU.
  - Create a Python script to set up a Docker instance that serves the model. This will ensure the environment is consistent and portable.
- Inference:
  - Develop a script to perform inference the HumanEval dataset.
  - This script should interact with the served model to generate predictions for the provided samples.
- Evaluation:
  - Use a sandbox environment (docker instance) to assess the pass rate of the HumanEval results obtained from the Qwen/Qwen2.5-Coder-0.5B model.
  - This evaluation will help determine the effectiveness of the model's predictions. We expect pass@1 > 0.5, please tune your prompt and response post-processing to achieve that.
- Performance & Quality Improvement
  - How can you improve the HumanEval's metric? Be open-minded.
  - How can you enhance the performance of the inference and evaluation processes.
  - How can you scale this evaluation process and make it run faster?

Evaluation Criteria

1. Evaluator must be able to clone repo and replicate the same setup and eval to verify claimed pass@1 percentage.
2. Quality of code and prompts.
3. Thought process and idea quality in answers to 'perf and quality improvement' Qs