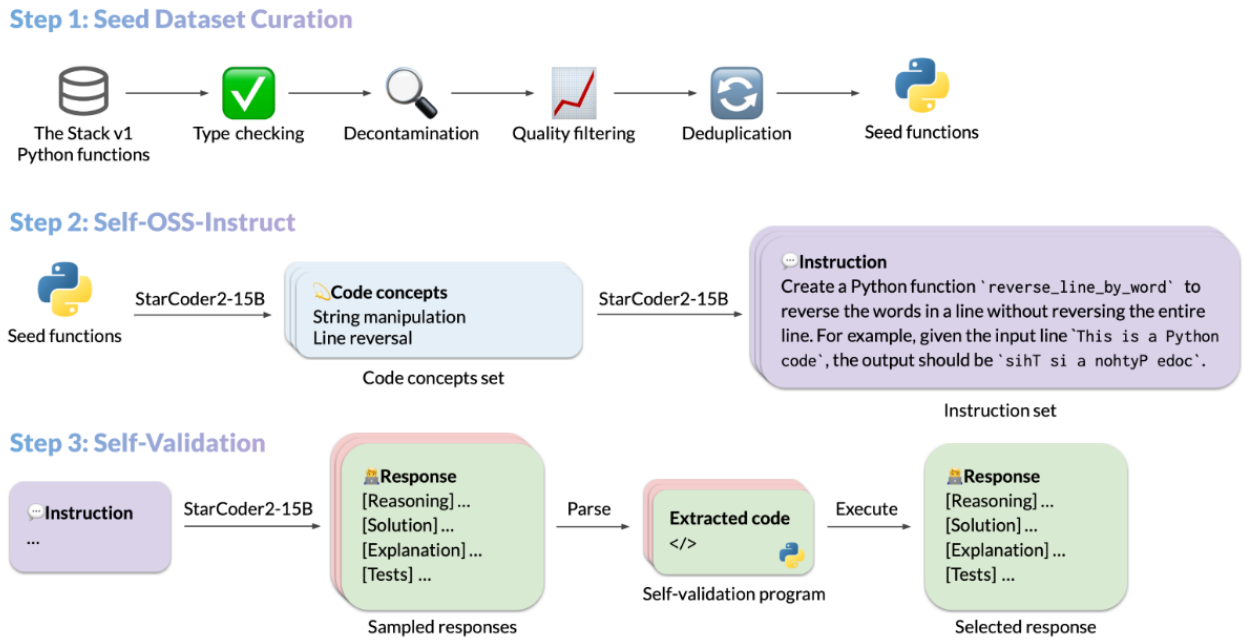
Link to the GitHub: https://github.com/bigcode-project/starcoder2-self-align/tree/main

Goal: This pipeline helps to generate thousands of instruction-response pairs. Our goal is have a dataset which is in the format instruction-response pairs, which can be used in finetuning instruction models.

Model: https://huggingface.co/bigcode/starcoder2-15b

Datasets: https://huggingface.co/datasets/bigcode/the-stack-v2

Overview of the framework:



Note:

- The code on GitHub is for the stack v1. You have to work on the stack v2.
- You have to choose programming languages: Java, C++, C#
- For the step 1: https://github.com/bigcode-project/starcoder2-self-align/tree/main/seed_gathering
- For the step 2 and 3: https://github.com/bigcode-project/starcoder2-self-align/tree/main/src/star_align

Step1: Seed Gathering

I attached SeedGathering.ipynb in the email. You can follow the instructions here: Seed Gathering (Step1). Please focus on the ***download_contents*** function in the notebook file to get the content of the data point (different from the original code). There are 3 sub-steps for Seed Gathering, make sure that you save the data after each sub-step. After the last sub-step, you should rename the column **"content"** to **"seed"** before saving.

Step 2 and 3: Self-OSS-Instruct and Self-Validation

You can follow the data generation pipeline on Pipeline. However, we can use vLLM to open the vLLM server instead of docker. You can follow the instruction on this documentation: vLLM. After that you should set the **environment variables (OPENAI_API_KEY, OPENAI_BASE_URL)** that match your setting. After that, you can follow the instructions on the pipeline. Please do it step by step, "S->C" -> "C->I" -> "I->R". Make sure that you save the dataset after each sub-step.

Notes:

- This pipeline is built for Python. You should pick another programming languages such as Java, JavaScript, C#,... for this project. The idea is similar, but you need to use a different parser that match the language that you choose. Also, you have to re-format the prompt to match your programming language.
- The code on GitHub is for the stack v1. You have to work on the stack v2. You may need to change the code a little bit to match the format of v2.