Step 1: Data Collection

Unsafe Prompts

Beavertails

Safe Prompts

Alpaca

Jailbreak Prompts

Do-Not-Answer, CatQA, AdvBench, Forbidden Questions

Step 2

Multilingual Translation Using Google Translate API

SEALSBench
A Multiligual Safety Benchmark

Training

Validation

Testing