



## Step 1: Data Collection



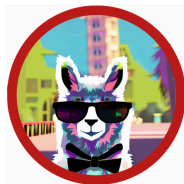
Unsafe Prompts

Safe Prompts

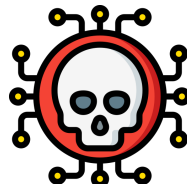
Jailbreak Prompts



Beavertails



Alpaca



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

## Step 2



Multilingual Translation Using  
Google Translate API

## SEALSBench

A Multilingual Safety Benchmark



Training



Validation



Testing