CONFIDENTIAL INFORMATION

This document contains confidential and proprietary information intended solely for the use of the individual or entity to whom it is disclosed.

The aim of this project is to **gather data** on how **AI models should handle interactions** that deal with **sensitive topics** or **harmful themes**.

🙌 **By participating in these projects, you are actively contributing to ensuring the safety of cutting-edge LLM models across critical dimensions.** 🙌

Handbook Navigation

⭐ Project Overview

Welcome to the **Pangolin Safety** projects! These projects generate critical data that is fundamental to ensuring AI safety. This group of projects include: prompt generation (text & images), response ratings, and response rewriting. All of these different interactions train the AI model to understand how it should handle sensitive or harmful topics.

Here are some key definitions to help you contextualize the work on these projects:

- **SFT (Supervised Fine-Tuning):** training the model by giving the model "perfect" prompt-response pairs to demonstrate how it should respond to different requests via examples.
- **RLHF (Reinforce Learning from Human Feedback):** training the model by giving it direct feedback about how it is performing on specific criteria via ratings.
- **Vision:** in real life, these models encounter various data types (i.e., text, image, video, audio, etc.), so we need to train the model to be able to handle all of them.
- **Turn:** a prompt-response pair (i.e., you make a request to the model and it responds to the request creates this prompt-response pair).
    - **Single-Turn:** one user prompt and one model response to that prompt.
    - **Multi-Turn:** a continued back-and-forth interaction between a user and a model where there are multiple prompts and responses.

Important Notes on Sensitive Content and Contributor Wellness

Sensitive Content: This work may involve reading or generating sensitive content (e.g., swearing, violence, slurs). You control what you request **and may opt out at any time**.

Sensitive Content Wellness Support:

Group Sessions. The project offers weekly group wellness sessions for contributors. There is a dedicated facilitator in these sessions to help navigate the discussions across various

wellness topics, including meditation, maintaining work-life balance, and mindfulness. These sessions may also be tailored to the type of content or challenges specific to project groups.

1-on-1 Sessions. The project offers private 1:1 sessions for contributors with a licensed wellness practitioner. These sessions allow contributors to discuss any topics relevant to their wellbeing and are completely confidential.

Sensitive Example Prompts: Some example prompts may be configured in a way that is inherently objectionable or offensive, for example allowing explicit content but disallowing either LGBT or heterosexual. The example prompts never reflect how we think the AI should behave, but rather test whether it is customizable to a given ruleset.

*Note: These **instructions** include examples with **swearing**.*

❗⚠️ Critical Project Restrictions ⚠️❗

Do not ask the AI to locate or discuss **Child Sexual Abuse Material (CSAM).**
Do not ask the AI to generate text involving **sexual content with minors (i.e. under 18)**.

Violation of these policies will result in disablement from this project.


Thanks for your help ensuring AI safety!