**Multilingual Centralized Screening Module**

**Welcome to the Multilingual Centralized Screening Module! This module is designed to prepare you for the screening by 1) providing a refresher on the screening instructions you were introduced to before and 2) give you a chance to work through a practice task. Let's get started!**

**Resources:**

[**Multilingual Centralized Screening Instructions**](https://docs.google.com/document/d/e/2PACX-1vQVtVG-2xqjCOmD2l8Gynuh0T8BUK3R_mo4XB-Vu48UR5hHEBN5cEFyW1jUcRhOuufhF0T5F6okYMSr/pub)

Video: 1

Cypher RLHF training task walk through

Step by step completing task in this project

* Check the language requirements for the task
* Check task types like summarization, classification or rewrite.
* Familiarize yourself with the goal and requirements. For the specific task type working on. Keep in mind that the goal is to write a prompt that causes a model failure.
* You will need to review the prompt requirement. Your goal is to write a new prompt based on task types if it requires reference text u might need to review those guidelines as well.
* Your reference text should be copied and pasted into the box provided
* Refer to the instructions linked at the beginning of this course for steps to insert the reference text.
* Post insert or no insert you will indicate the prompt category
* Necessary checks for instructions following and truthfulness
* You will need to verify at least one of the model responses produced a failure before rating the responses.
* After verifying a response failure we have to evaluate both response individually based on 5 key dimensions instructions following , truthfulness ,conciseness ,writing quality ,harmlessness and localization. Use rubric to access
* Look for issues like factual in accuracies , failure to follow instructions or in appropriate tone.
* You can toggle between each response A or B on the left side of the screen,
* Then we will rate response using the scale 1 to 3
* After rating we need to rate over all response quality on scale 1 to 5 based on rubric.