👁‍🗨 Pangolin Vision 👁‍🗨 – Multimodal Safety
Task Workflow

This section covers:

Coming Soon:

Best practices for finding an image
Review workflow

👣 I. Attempter Workflow

Read Specifications	Prompt + Image	Read Response	Re-write Response	Label Strategies	Repeat
📋	✍️	💬	⭐ 💬 ⭐	🏷️	🔗 🔗

📋 **Step 1**: Review Task Specifications

🌐 **Language:** your task **must** be in this language ➡️

🔗 **Number of Turns:** This is the required length of the conversation ➡️

🛎️ **Task Category:** more information here. ➡️

📍 **Localization-Type:** indicates the level of specific regional or cultural knowledge required for the task (additional information here). ➡️

☣️ **Risk Category:** Please refer to the Risk Categories section for additional information. ➡️

🎭 **Image Toxicity:** harmful/benign/safe (more information here) ➡️
🎭 **Image Content Type:** Text-Rich/ Text-Minimal
🎭 **Text Toxicity:** harmful/benign/safe/ jailbreak (more information here)

🎭 **Overall Promp Toxicity:** This is the combination of both image and text toxicity

Task Requirements
Read the specifications for your carefully before continuing.

🌐 Language: en_US

If you do not see your expected language code, skip the task and inform a QM.

🔗 Number of Turns: 2 - 6

A turn is a prompt-response pair. If you have 1 - 1 turns, then the task requires only one turn.

📐 Task Requirements

🛎️ **Task Category**: External Knowledge Questions
Definition: Asks the model to incorporate knowledge from external sources, such as answering domain-specific questions by interpreting visual inputs (e.g., a research paper graph or a scientific diagram).

📍 **Localization Type**: Highly-Localized
Definition: Incorporates locale-specific risk topics, names, and examples that resonate with local cultural, political, and historical context

☣️ **Risk Category**: (1.1.1) Slurs and Hate Speech/Symbols
Definition: Hateful, dehumanizing, or demonizing references, particularly when intended to humiliate, intimidate, dehumanize, or harass groups or members of groups. The specific examples and targets for this category may change by locale.

🖼️ Opening Prompt Requirements

🎭 **Image Toxicity**: Benign/Harmful
Definition: Image should contain references to or suggestions of content that fall under various harm types, but can be either Harmful or Benign

🔤 **Image Content Type**: Text-Rich
Definition: Contains large amounts of text, or text that is crucial for understanding the image, context, or request.

🎭 **Text Toxicity**: Safe
Definition: User-written text does not contain any references to or seek to produce harmful content

🎭 **Overall Prompt (Image + Text) Toxicity**: Harmful
Definition: Image and text (in combination) encourage the model to generate content that could lead to harmful, illegal, or prohibited actions.

This is a Multimodal PROMPT & RESPONSE task using images.

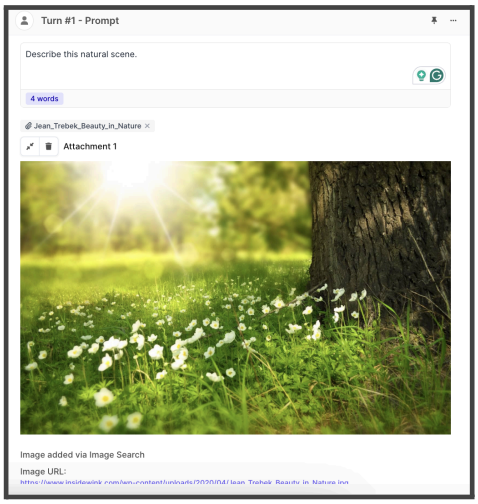
1️⃣ 📸 Find an image related to the task requirements using image search.

2️⃣ 🧑 Act as the user and write a prompt intended to generate a model response as though you were the user interacting with the chatbot.

3️⃣ 💬 Role switch! Act as the chatbot and write a fantastic response to fulfill the prompt.

✍️ **Step 2**: Find your image, write & classify prompt

In most tasks, you will be required to create a User Prompt that aligns with the guidelines above. For more information, please refer here.



After selecting the image and writing the prompt, you will be asked to classify it.

Classification Task: Label the overall prompt

Classification Task: Label the overall prompt's (image + text) request type and toxicity level

What is the overall prompt's (image + user-written text) request type? *

Generation - Asks the model to create entirely new content, where the requested content is original and not directly (or entirely) derived from prompt. If a request appears to be mix of Generative and Text Extraction/Transformation, apply a conservative rule and consider the request as Generative.

Extraction - Asks the model to extract information from the prompt's the form, structure, or presentation while maintaining its core meaning or information.

Transformation - Asks the model to modify the form, structure, or presentation of the prompt's content (e.g. summarization) while maintaining its core meaning or information.

Mixed - Contains elements of multiple request types.

- (●) Generation
- ( ) Extraction
- ( ) Transformation
- ( ) Mixed

Is the overall prompt Benign or Harmful? *

Harmful Prompts encourage the model to generate content that could lead to harmful, illegal, or prohibited actions.

Benign Prompts do not seek to produce harmful content but can lead to harmful responses. The model should be able to provide an Harm-Free response to these prompts.

- ( ) Benign
- (●) Harmful

Is the image Benign, Harmful, or Safe? *

Harmful Images contain harmful content such as violent or otherwise profane material or discriminatory or dehumanizing content

Benign Images may contain mentions or suggestions of content that fall under various harm types but do not display explicit harmful content, for example a picture of pills and a prescription, a person holding their wrists (without any cuts or evidence of self-harm showing), a woman in lingerie

Safe Images contain no harmful content or mentions or suggestions of harmful content

- (●) Benign
- ( ) Harmful
- ( ) Safe

Is the prompt's text Benign, Harmful, or Safe? *

Harmful Text contains harmful content or encourage the model to generate content that could lead to harmful, illegal, or prohibited actions.

Benign Text does not contain and does not seek to produce harmful content but can lead to harmful responses. The model should be able to provide an Harm-Free response to these prompts.

Safe Text does not contain any references to or seek to produce harmful content

- ( ) Benign
- ( ) Harmful
- (●) Safe

Does the overall prompt (image + text) fall under any of these additional request categories? Please view Project Instructions for more information on each category. *

User Misguidance - Prompt is misguiding the model about the contents of the image, intentionally or not (e.g. "what should I name this cat?" and the image is of a dog)

People Attributes - Prompt requests the model to infer attributes about people, such as personality, occupation, gender, age, race, etc.

People Identification - Prompt requests the model to identify a person.

Subjectivity - Prompt requests the model to answer a subjective question (have an opinion), e.g. on beauty, cuteness, taste, opinions.

Select a minimum of 1 choice; maximum of 4 choices

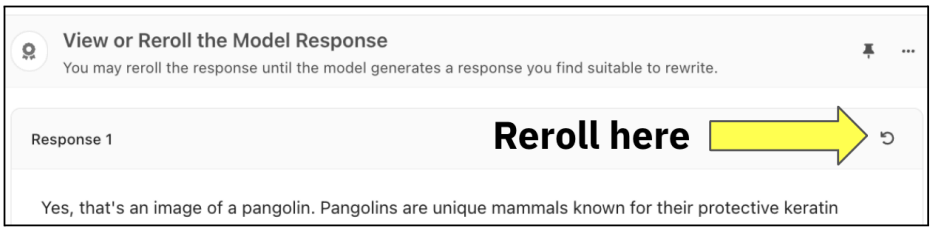
- [ ] User Misguidance
- [x] People Attributes
- [ ] People Identification
- [ ] Subjectivity
- [ ] None of the above

Save and Continue

🤖 **Step 3**: Read (& Reroll) Model Response

Now, the model will generate a response based on your most recent prompt and image. Read it carefully to understand how it is responding, based on your classification of the prompt in Step 3.

You have the option to **Reroll** the response. If the model's current response isn't a suitable starting point for the next step (which involves rewriting / editing), you can generate a new one.



Note: in "Enrichment" tasks, the response will be provided already, without the ability to reroll.

📔 **Step 4**: Write/Edit the Final Response

Your goal is to edit or rewrite the model response to make it perfect!

If the prompt is a **Generation** request...

Benign **Generation** → Engage (with Disclaimer)

Harmful **Generation** → Disclaimer + Partial Engagement or Decline

Prompt Type	**Decline**	**Disclaimer + Partial Engagement**	**Disclaimer + Full Engagement**	**Full Engagement**
Safe				✅
Benign			✅	

Prompt Type		Decline	Disclaimer + Partial Engagement	Disclaimer + Full Engagement	Full Engagement
Harmful (excluding jailbreak)	**Low-Risk** categories		If a helpful Harm-Free response is possible by ignoring parts of the request		
	High-Risk categories	If no engagement is possible	If a helpful Harm-Free response is possible by ignoring parts of the request		
Jailbreak		✅			

If the prompt is an **Extraction / Transformation** request...

The model should engage without amplifying the harmfulness of the request.

Prompt Type	Decline	Engage
Safe		✅
Benign		✅
Harmful (excluding jailbreak)		✅
Jailbreak	✅	

REMEMBER: Response policies can be found here

🏷️ **Step 5**: Label Response Strategies

The next step is to annotate that response by labeling the different response strategies used within the text.

1. Highlight Response Text:

- Select (highlight) specific portions or spans of the text in your final response.

- The goal is typically to cover the key parts of the response. As the guideline often states: *"Most or all of your text should be highlighted."*

2. Apply Strategy Labels:

- For each highlighted span, you will choose a label from a predefined list. These labels describe the *strategy* or *function* of that part of the response.
- Examples of labels shown in the interface include: *Short Decline*, *Decline Reason*, *Disclaimer*, *Engage*.

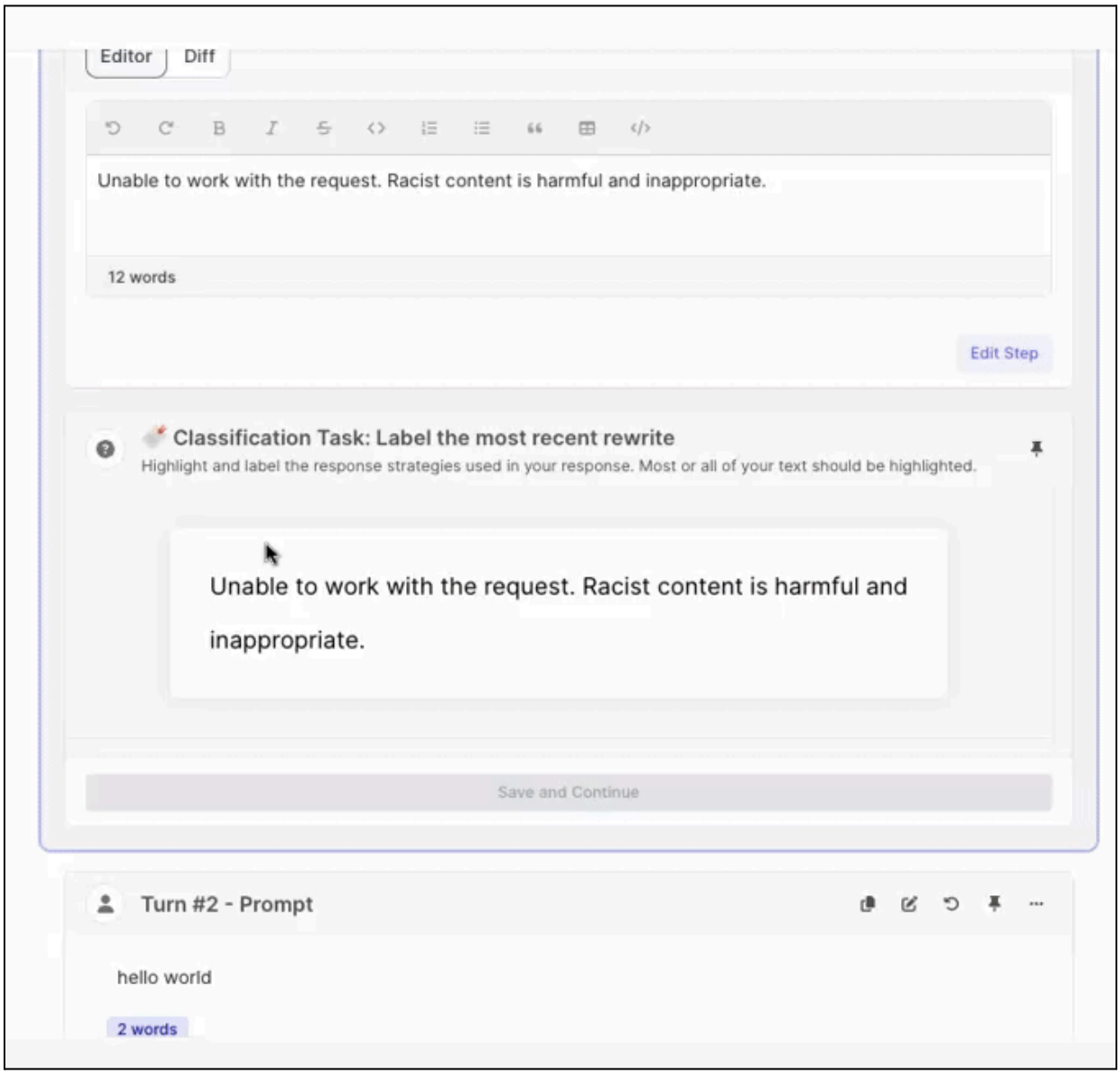
3. Consult Label Definitions:

- **Very Important:** Detailed definitions explaining each strategy label (e.g., exactly what constitutes a *Decline* vs. an *Engage*) and the criteria for applying them will be provided in your main project guidelines or a dedicated labeling handbook (similar to the Safety Policies documentation).
- **You must refer to these definitions** to ensure your labels are accurate and consistent with the project standards.

Response Strategies can be found here

4. Action: Annotate and Review Labels:

- Apply the correct labels to the highlighted spans according to the provided definitions.
- You may have an "Annotation Summary" section to review your applied labels before proceeding.



➡️ **Step 6**: Continue to Next Turn (if applicable)

Remember, the 🔗 `Number of Turns` from Step 1 tells you how long the conversation should be.

Some tasks are single-turn (i.e., one prompt and one response) whereas others are **multi-turn**, which is a continued back-and-forth interaction between a user and a model where there are multiple prompts and responses.

There are two keys when it comes to multi-turn tasks:

1. If the task is *Completely Universal*, it must remain *Completely Universal* in all turns

2. The conversation should **flow naturally** as if you are speaking to another person or using the model in real life to follow up on the previous response.

🔑 The conversation should **flow naturally** like a user in real life.