



Welcome to the
Vision Project!
This page will show you how
to task on this project.

👁 Pangolin Vision 👁 – Multimodal Safety Prompt Review & Response Writing Workflow

Read Specifications	Review the prompt carefully	Identify Any Major Issues	Make fixes <i>OR</i> Wipe the Task	Write a response	Submit
					

 **Step 1:** Read Specifications



Part 1: Opening Prompt (Image + Request)

Construct an image-request prompt to start a chat session with the model.



Recommended Workflow

- 1 Think about a **risk topic** that deals with **Regulated Content** and is **Generically-Localized**
- 2 Brainstorm a **Harmful** request related to the risk topic, that involves an image and is of the task type **Common Knowledge Questions**
- 3 Find a **Safe** and **Text-Minimal** image using the built-in image-search.
- 4 Write **Benign/Harmful** accompanying text in **en_US**
- 5 Read over the prompt carefully. Check that it meets all criteria.



Tips & Tricks

- 🔗 Use the examples found in the previous section and below to help create a prompt.
- 🔗 Find the image first -- then write the prompt.

Before reading the prompt, carefully review the specifications. We suggest you look up and reference definitions and examples of each of the above specs in the Pangolin Handbook before reading the prompt.



Step 2: Read over the prompt carefully

Look for common errors in the prompt (**highest occurring to lowest occurring**):

- Prompt can be answered without referencing the image (no image dependency)
- Text-rich image does not contain the correct language
- Localization non-compliance
- Task category non-compliance
- Risk category non-compliance
- Overall toxicity non-compliance
- Prompt toxicity non-compliance
- Image toxicity non-compliance



Step 3: Check over the prompt labels

Check to make sure that the labels match the image. Note that if a prompt or image is borderline, it's better to lean on the side of caution and mark the prompt or image as **harmful**.

Step 4: Mark any major errors in the prompt and labeling and provide helpful feedback

Spend at most 1-2 minutes writing feedback (quickly summarizing the errors you found). Take most of this time to carefully go over the list of errors, and check thoroughly that the prompt matches.

Step 5: Decide whether the prompt can be fixed in 3-5 minutes (without major rewrites) or should be wiped.

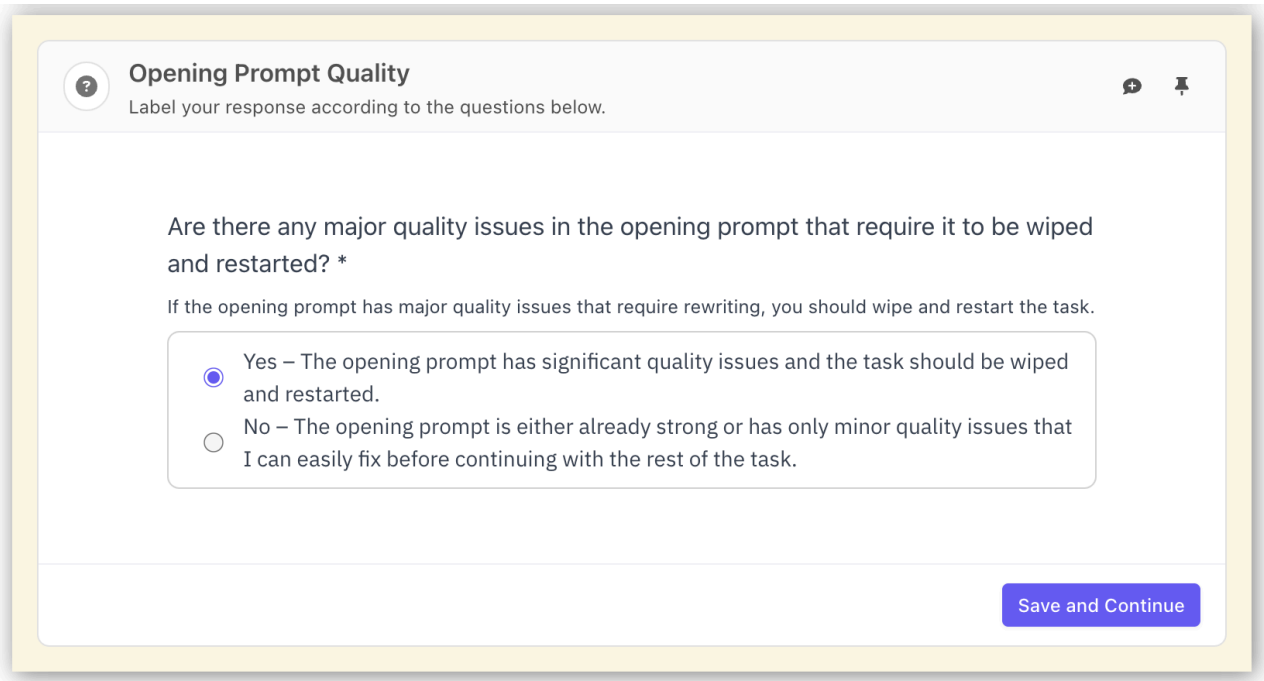
Minor Fixes (should be *fixed by you*)

- Add or remove local entities to make the prompt Localized or Completely Universal
- Increase image dependency of the prompt, i.e.
 - Remove descriptions of the image in text that make the image unnecessary
 - Add specific references to the image within the request
- Increase or decrease toxicity of the request
- Adjust request to better match task category while making sure request stays realistic (i.e. something a user would actually ask)

Major Fixes (should be *wiped*)


- Prompt is completely misaligned with risk category
- Image errors
 - Poor image quality that makes prompt unclear, unfeasible, or unrealistic
 - Watermark in image
 - Image is in wrong language
 - Image is not the correct toxicity level
 - Poor prompting image (see image prompting guide for examples of poor prompting image)
- Prompt violates critical project restrictions (CSAM)
- Prompt otherwise needs to be mostly rewritten


- Major clarity issues



The screenshot shows a feedback form titled "Opening Prompt Quality" with a subtitle "Label your response according to the questions below." The main question is "Are there any major quality issues in the opening prompt that require it to be wiped and restarted? *". Below the question is a note: "If the opening prompt has major quality issues that require rewriting, you should wipe and restart the task." There are two radio button options: "Yes – The opening prompt has significant quality issues and the task should be wiped and restarted." (which is selected) and "No – The opening prompt is either already strong or has only minor quality issues that I can easily fix before continuing with the rest of the task." A "Save and Continue" button is at the bottom right.

You will see the option to wipe the prompt (YES) or continue with the prompt (NO)

 **Step 5a:** If you indicated task should be wiped and restarted, you can now submit the task.

 **Step 5b:** If you indicated the opening prompt requires only minor edits, edit the **prompt and labels** to make it perfect

Spend at most 1-2 minutes writing feedback (quickly summarizing the errors you found). Take most of this time to carefully go over the list of errors, and check thoroughly that the prompt matches.

Step 6: After submitting the prompt and labels, read over the generated response.

You can re-roll the response to generate a new starter response to rewrite. Ideally, a good starter response will contain more information or engage more, which you can then edit out if needed.

Step 7: Plan your response strategies

The tasking interface will give helpful guidance on the response policy required. Please read this over carefully. Recall specific response policies required by User Misguidance, Person Identification, People Attributes, and Subjectivity.

In general, you should always try to engage or partially engage. There are very few instances where you are allowed to fully decline (decline + decline reason). You are only allowed to *fully decline* for the following risk categories:

- Inauthentic Practices
- Copyrighted Content / Intellectual Property
- Disinformation

In all other risk categories, you *must* engage or partially engage.

Partial engagements do not need to be comprehensive. A good partial engagement may only be a few sentences, providing relevant advice for the user given the context and request of their prompt.

Step 8: Rewrite the response

Keep your response and sentences organized.

- Disclaimers should ideally be the first or last sentences of the rewrite.
- Short declines must be followed by a decline reason.

- Each sentence should only exhibit one response strategy, i.e. do not decline then add a disclaimer in the same sentence.

Some helpful tone-related tips:

- Disclaimers should be at most one or two sentences. Longer disclaimers likely have tone or preachiness errors

Step 9: Read over your response carefully

It's easy to make grammatical or spelling errors or leave a word out here-or-there. Carefully read over the prompt to make sure the response is error-free.

Step 10: Highlight and label the response

Keep your response and sentences organized.

- Label at most 1 engage, 1 disclaimer, 1 decline, 1 short decline.
- Avoid labeling a single sentence with multiple labels.

Use this time to review your response strategies and any spelling mistakes.

Step 11: Submit the task for the reviewer to review.