Publicada con documentos de Google Denunciar uso inadecuado Más información

Cypher RLHF Revamp Instructions Doc Actualizado automáticamente cada 5 minutos

Cypher_RLHF

General Instructions Revamp RLHF

Attempter Specifications

✍️Updates/Changelog
Project Overview
Task Attempt Workflow
Task Overview - This is a single turn project
  Step 1 - Writing Prompts
    Prompt Writing process
    What Qualifies as a Response Difference?
    What is a Contrived Prompt?
    Type of Prompts
      Common Errors with Prompt Categories
  Step 2 - Rating Model Responses
    1. Localization
    2. Instruction Following
    3. Truthfulness
      3a. Truthfulness v/s Instruction Following - What's the deal? Save this chart
    4. Response Length
    What's the deal with pleasantries?
    5. Structure, Writing Style, and Tone
    6. Harmlessness/Safety
      When Model Claims to be Human or Have Emotions or Opinions:
    7. Overall Satisfaction Score
  Step 3 - Preference Ranking
    Writing Justifications for Preference Ranking
  Step 4 - Response Rewrites
Appendix
  Confusion between Dimensions
    Mixed Language Guidelines

✍️Updates/Changelog

- 4/22 - Additional details on major vs. minor errors for Truthfulness and Instruction Following. Also added a note on LaTeX in Structure, Writing Style, & Tone.
- 3/25 - Nuances added to the “Truthfulness” rating dimension, check it out here!
- 3/20 - Added in the instructions this info in Instruction Following and Response Length sections of the instructions:
  - ⚠️**Note:** The client has clarified that if a response fails to meet explicit length constraints in the prompt, it is by definition also at least a minor issue for Response Length evaluation.

Project Overview

Welcome to Cypher RLHF. You'll guide the AI model through natural, real-world conversations, ensuring its responses are accurate, relevant, and reflect how users interact in everyday scenarios.

Your input is vital to refining the AI's ability to learn from real interactions. We're excited to have you involved in this project as we strive to enhance the model's capabilities.

Task Attempt Workflow

Here are the steps in a Cypher task:

- Write prompt
- Rate Model Responses
- Rank Your Preference
- Rewrite Response

IMPORTANT! ChatGPT or other AI tools are NOT PERMITTED to create prompts or evaluate responses. Using AI tools will result in a flag on your account for removal from the project and can lead to eventual removal from the platform

- On the right hand side of the task, you will be provided with a prompt category, please create a prompt based on this prompt category.
- The goal is to **write unique, creative, diverse prompts that adhere to the prompt categories. The prompts should be natural, reasonable prompts for a real user to ask.** In doing so we aim to have **tangible differences between two model responses.**
- **Prompts should not ask the model for information that requires knowledge after April 30, 2024.**
- **Some prompts may need a reference text, some prompts may not need a reference text. Reference text should be in target language.**
- **Either the prompt OR reference text must be localized.** Localization means the model's response feels authentic and relevant to a specific region or country. This involves:
    - **Reflecting Local Knowledge:** Knowing key locations, customs, and events.
    - **Understanding Regional Nuances:** Adapting references to suit cultural norms.
    - **Meeting Specific Needs:** Addressing local laws, regulations, or processes.
    - **Tailoring to Cultural Context**: Ensuring the response is closely adapted to the cultural background of the target audience.

Step 1.5: Iterating on a prompt

- When generated, prompts **must** produce two responses with clear **differences**. The differences can be quality based (e.g. one makes a mistake), or more subjective (one is more thorough or uses different structure). **We should NOT force model mistakes at the expense of a natural prompt.**
    - **PLEASE NOTE**: This does **NOT** mean you should create arbitrary and unnecessary constraints, such as "don't use the letter H" or "bold every other word." We often see attempts to make prompts more complex in the hopes of "confusing" the model, when in reality it just produces poor data. Check out our **contrived prompts** section.

Step 1.75: Inserting the Reference Text

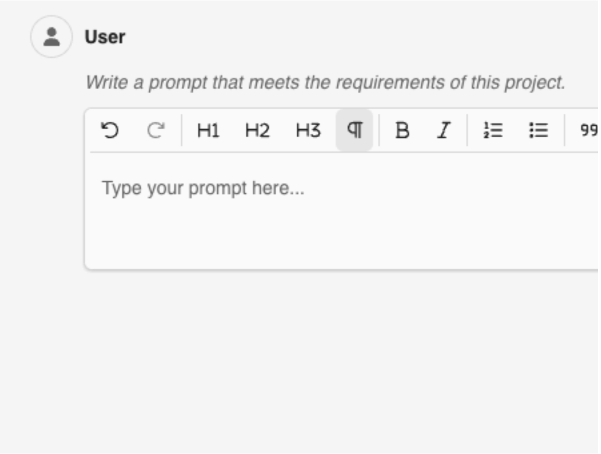
If you see a clip icon, add the reference text in the box after clicking the clip icon. **Please ensure that the reference text is entered in the designated input box that appears.**

If you don't see a clip icon, the prompt category does not require a reference text.

Reference text should be from credible/reputable sources. You can source this information from Online Articles and Websites, News Outlets, Specialized Blogs and Industry Sites, Books and E-Books, Academic Journals and Papers, Government and Official Publications, Open Data Repositories, Educational Platforms and Courses, Multimedia Transcripts, Public Domain Text, etc.

A reference text should be between 200 and 400 words. For Japanese and Chinese locales, we should aim for 300-600 characters.

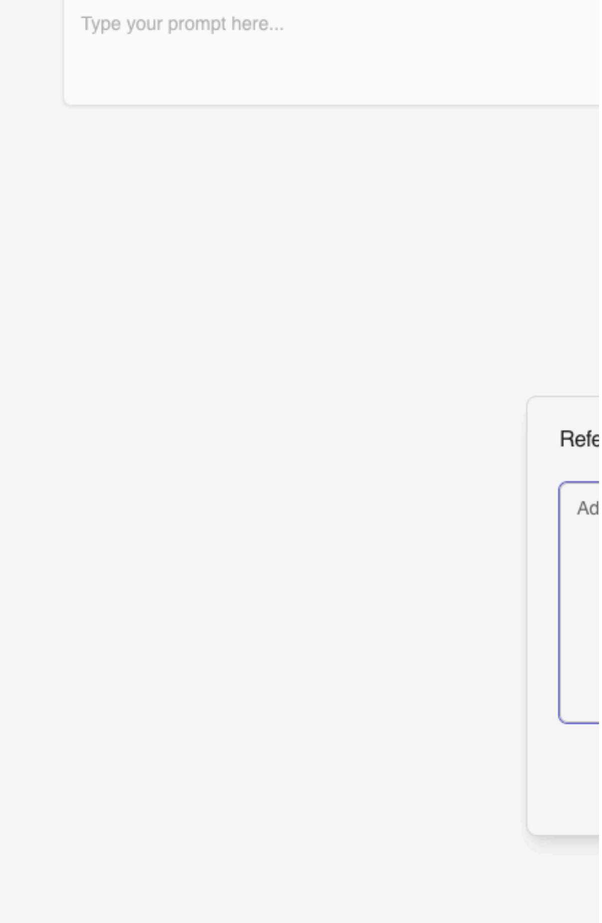
Clip Icon:



User

Write a prompt that meets the requirements of this project.

Publicada con documentos de Google Denunciar uso inadecuado Más información



Step 2: Rating each model response according to the dimensions

- There are **7 dimensions** including an overall quality score and some direct questions about the content of the response.

- The dimensions are:
    - **Localization:** Does the response use correct spelling, grammar, and word choice for the region? Does it look like it's coming from a local from your locale?
    - **Instruction Following:** Does the response do what the contributor asks? Does it meet the explicit and implicit requirements of the prompt?
    - **Truthfulness:** Is the information factually/contextually accurate? - are there any inaccurate statements or misrepresentations of the information?
    - **Response Length**: Does the response efficiently deliver only the necessary information in a clear, concise, and structured manner without unnecessary repetition or filler? Look for length, relevance and repetition.
    - **Structure, Writing Style, and Tone:** How well the response reads: does it have good structure, appropriate tone, clarity etc.
    - **Harmlessness**: The extent to which the written response avoids harmful elements such as racism, offensive language, or any content that may cause harm or discomfort to others
    - **Overall Satisfaction:** When rating the overall satisfaction,

Publicada con documentos de Google Denunciar uso inadecuado Más información

that fulfill the user's request effectively, even with minor errors, should be graded more generously, while those that lack obvious errors but fail to feel satisfying should be slightly penalized.

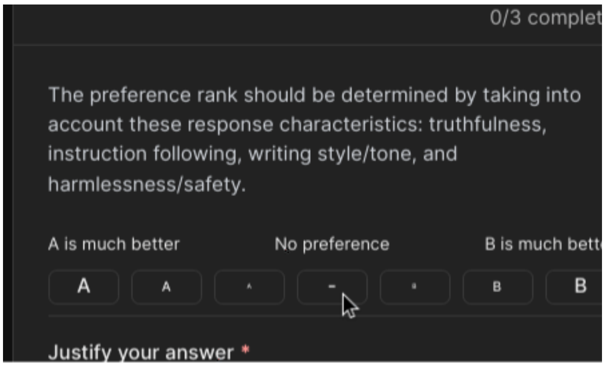
- Rate each dimension according to:
    - **1 - Major Issues**
    - **2 - Minor Issues**
    - **3 - No Issues**

- Some dimensions will require a short justification:
    - Stay **brief** and **specific**, only mention what is incorrect.
- Do an overall score of the response:
    - Take all your individual ratings for the dimensions and provide an overall score for the response.

***Step 3*: Preference ranking on a 1-7 Likert Scale**

- Select 1 or 7 if one response is **much better** than the other response:
    - One response has no key issues while the other has major issues.
    - One response is significantly more satisfying than the other.

- Select 2,3 or 5,6 if one response is **better or slightly better** than the other:
    - Only relatively minor or subjective differences between the responses.
    - One response is somewhat more satisfying than the other.



The preference rank should be determined by taking into account these response characteristics: truthfulness, instruction following, writing style/tone, and harmlessness/safety.

Justify your answer *

***Step 3.5*: Writing a justification to explain the preference ranking**

- Look here for a full rundown on the requirements of a good justification.

- A good justification should:
    - Provide **sufficient evidence** from the responses to explain the preference ranking.
    - Follow these 4 requirements:
        - Starts with a clear statement of the **verdict.**
        - Mentions the **1-2 most critical reasons** for the ranking with **specific** evidence from **each response**.
        - Is typically **2-3 sentences**.
        - Is edited to trim down any **unnecessary** information or words.

- If there are multiple issues differentiating the quality of the 2 responses then the focus should be on the most prominent ones.

***Step 4*: Rewriting the selected response**

- Look here for a full rundown of how to do response rewrites.

MAX.

- Rewrite Step by Step Guide:
    - Read the prompt again.
    - Fix any critical issues you identified in the prior response rating. Again, please do not change more than you need to. Some examples of fixes include:
        - Correcting inaccurate statements.
        - Ensuring the response addresses all aspects/constraints of the prompt.
        - Fixing spelling, grammar, and formatting mistakes.
    - Re-read the prompt.
    - Double check all issues are resolved and new issues have not been introduced.

Step 1 - Writing Prompts

Prompt Writing process

On the right hand of the task, you will be provided with a Prompt Category and instructions on how to write prompts for that category. Ensure that your prompt aligns with the correct category (**this is mandatory**) and select the appropriate option in the multiple-choice question that asks which category your prompt falls under.

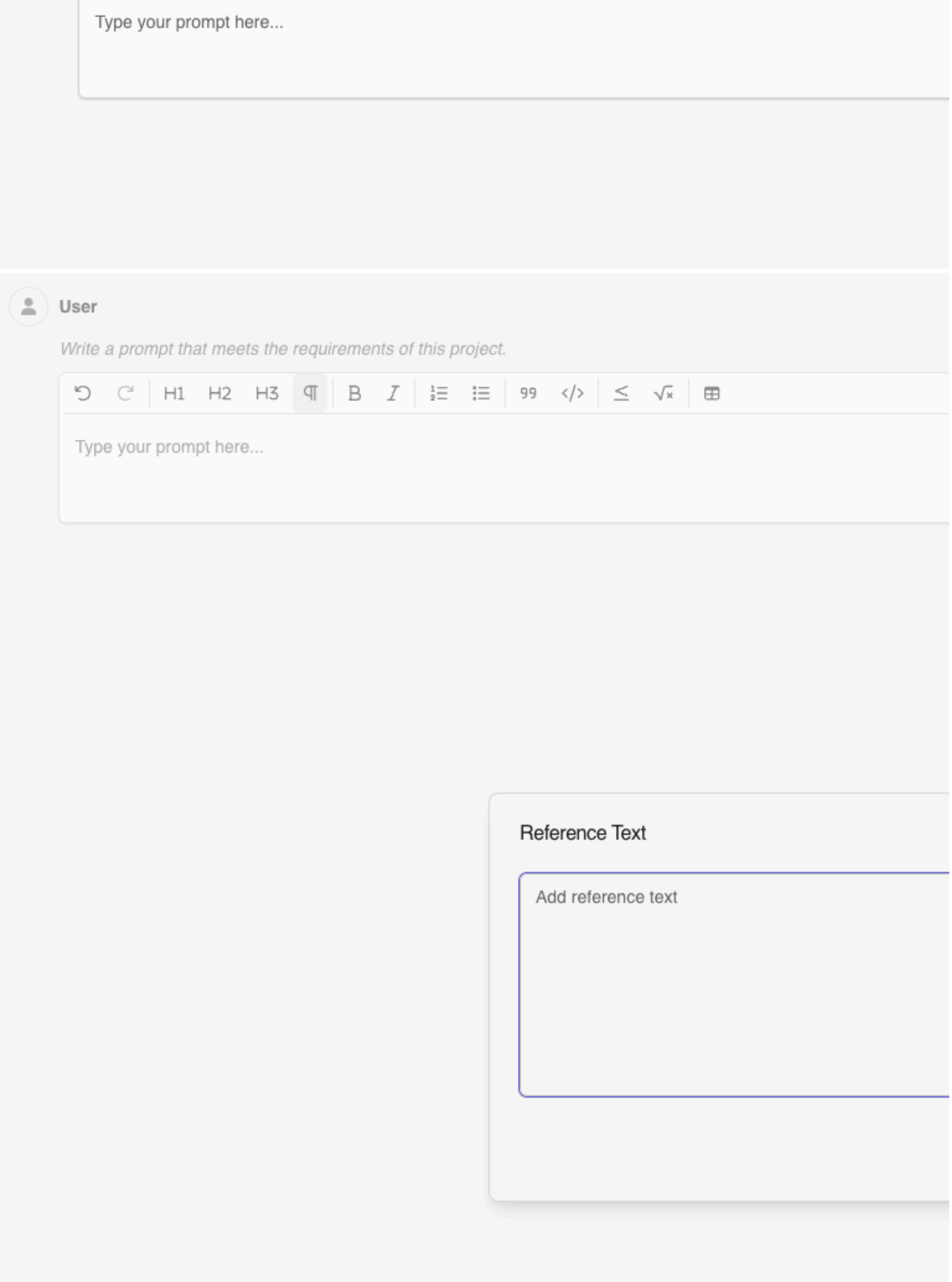
What is a Reference Text?

Some of these tasks will need a **Reference Tex**t. Reference text is a short article or body of text that you need to provide along with the prompt. Reference Text is used to evaluate the model's ability to comprehend the content and answer questions based on it.

Example:
Prompt: From the reference text, provide a list of 5 popular fruits.
Reference Text: Some popular fruits include apples, bananas, and oranges, which are widely enjoyed for their sweet flavors and versatility. Grapes, mangoes, pineapples, and cherries are also common choices, often eaten fresh or used in juices and desserts. Kiwi, peaches, and strawberries are favorites for their unique taste and texture. Watermelon, blueberries, raspberries, and blackberries are loved for their refreshing qualities, particularly in the summer months. Other fruits like papaya, pomegranate, avocado, lychee, and dragonfruit offer more exotic flavors. Additionally, pears, plums, apricots, and nectarines are frequently found in markets. Grapefruit, lemons, and limes provide a tangy contrast, often used in cooking or as a garnish. Finally, fruits like passionfruit, coconut, and cantaloupe add variety to the diverse selection of delicious and nutritious options available.

Where should you input the Reference Text?

If you see a clip icon, add the reference text in the box:



Type your prompt here...

User

Write a prompt that meets the requirements of this project.

Type your prompt here...

Reference Text

Add reference text

If you don’t see a clip icon, the prompt category does not require a reference text.

Where can I find the Reference Text?

Reference text should be from credible/reputable sources. You can source this information from Online Articles and Websites, News Outlets, Specialized Blogs and Industry Sites, Books and E-Books, Academic Journals and Papers, Government and Official Publications, Open Data Repositories, Educational Platforms and Courses, Multimedia Transcripts, Public Domain Texts, etc.

1. If the **prompt does not produce obvious differences between the two model outputs, try again**. The goal is to write a prompt which creates variations between model outputs, in any of our rating fields OR more subjective aspects.

2. **Either the prompt OR reference text must be localized.** Localization means the model’s response feels authentic and relevant to a specific region or country. This involves:

is closely adapted to the cultural background of the target audience.

3. Prompt writing should be an iterative process of fine tuning an input. While there are expectations for features of the prompt, the most important thing is to produce clear differences. Prompts can and should be rewritten several times to produce the desired result.

4. Things to keep in mind for Reference Text:
   a. Reference text should be directly related to the prompt.
   b. Reference text should be from credible/reputable sources.
   c. Reference text should be limited to only the relevant sections and should not be too long (minimum 200 to maximum 400 words). For Japanese and Chinese locales, we should aim for 300-600 characters.
   d. Reference text should provide sufficient detail and context for the prompt.
   e. Reference text should not require the model to access images or the web, as the model is not able to retrieve external content.
   f. Reference text should not use a URL exclusively as the source. The model cannot access content directly from URLs.

Note: When uploading or pasting the reference text, ensure that all information used to create your prompt is exclusively drawn from that text.

Important: do not use the same reference text in more than one task.

How do we know if the prompt is good?

The prompt is good if the two model responses are clearly different across content, formatting, tone, or dimension score (localization, instruction following, structure, harmlessness, truthfulness, or general satisfaction). The prompt is great if one of the two responses is clearly different from the other AND the prompt itself is not contrived.

A note on harmfulness and safety

1. **Avoid Harmful or Unsafe Content:**
   Do not write prompts that encourage harmful actions, illegal activities, or misleading information.
2. **Be Cautious with Sensitive Topics:**
   Avoid controversial or distressing subjects unless necessary. Ensure respectful and neutral language, free of hate speech or discrimination.
3. **Foster Ethical and Respectful Responses:**
   Write prompts that encourage positive, inclusive, and constructive responses, and avoid content that could harm, mislead or exploit others.

What Qualifies as a Response Difference?

The goal is to produce a difference between the two model responses in a non-contrived, natural way. A difference in response can include:

- **Content:** Both responses differ in the content provided to answer prompt.
- **Formatting:** Major difference in formatting between the two response.
- **Tone:** Tone of responses vary in a material way.
- **Dimension scores**: Differences in scores across Truthfulness, Instruction Following, Response Length and Writing Quality dimensions.
  - E.g.: Did 1 model response have at least 1 minor issue that the other model response did not?

Prompts should not ask for things that are **difficult to verify**:

- **Frequently Changing Information**: Prompts should not ask for something that changes frequently over time such as stock prices, or when books are published (since books are often released and re-released).
- **Information Past Knowledge Cut-off**: Prompts should not ask for anything that requires knowledge of events in 2024 (cut-off April 30, 2024).
- **Asking for links or for the model to read links**: The model cannot access links. It also cannot provide accurate links to

    letter 'K' in every sentence.
    - Explain the themes of '1984,' but make sure every sentence rhymes.
    - Explain how a smartphone works, but use only sentences with exactly ten words.

What is a Contrived Prompt?

A prompt is contrived when it imposes arbitrary, unnecessary, or overly specific constraints that do not contribute to the substantive understanding or exploration of the topic. A contrived prompt tries too hard to be diverse but is not original and ultimately is not helpful in a model's learning process.
For example:

- What might be the best easy way to get to Albuquerque, New Mexico from San Antonio, Texas. -> **Not Contrived Prompt**
- What might be the best easy way to get to Albuquerque, New Mexico from San Antonio, Texas Please make sure it is all in bullet points, each city mentioned is all bolded, the interstates are italicized and you explain it in 10 steps. -> **Contrived Prompt**
- What might be the best easy way to get to Albuquerque, New Mexico from San Antonio, Texas? My phone will be dead so tell me in a couple of steps so I can remember once it dies. -> **Not Contrived Prompt**

As you can see in these examples a prompt can differ in terms of its contrivedness, the point we make here is that often when trying to get the model response to fail we add unnecessary arbitrary details. Let’s avoid cases like the second point and shoot for diverse creative, natural, sounding prompts!

Note: Please do not use length restrictions unless integral to your natural ask of the model.

Type of Prompts

In the chart below, we cover the prompt categories for Cypher. Please be prepared to receive tasks from any of these categories. Please also check which of these require Reference Text. We want to stress that there is overlap between categories. For some prompts, Open QA can be similar to Brainstorming and Closed QA can be similar to Extraction and Classification. We want to explore the definitions and differences between the categories.

Prompt Type	Description	Examples	Has Reference Text?
Rewrite	**What it means:** You take a piece of writing and change how it sounds or reads, without changing its main point. **Imagine:** You have a paragraph written in a very formal style and you want to make it sound more casual. Or maybe you have something complicated and want it to sound simple so a younger kid can understand it. Unlike Summarization, you don’t necessarily make it shorter or focus only on the main idea; you’re just	• Rewrite this explanation in a way understandable to a 10 year old • Rewrite this text in the style of a haiku poem • Rewrite this text to be in active voice	**Yes**

Classification	**What it means:** You put things into categories or groups, usually based on some rules. **Imagine:** You have several paragraphs about the most common produce items in the United States and you separate them into "fruits" and "vegetables". **Difference:** Unlike Extraction, you're not just finding one piece of info; you're putting things into groups. And unlike Closed QA, you're not answering a question in a sentence—you're organizing data. ⚠️ **Note: Reference texts for classification should not be a standalone list of items, they should be articles, stories, and paragraphs containing the relevant items within. This is true for ALL categories but especially so for classification.**	• Categorize the following sentences into emotional categories: happy, sad, angry ◦ [include sentences to categorize] • Categorize the following nouns into cities or cars ◦ [include article containing cars and cities to be categorized]	Yes
Summarization	**What it means:** You take a longer text and shrink it down to its most important ideas, so it's shorter and easier to understand. **Imagine:** If you read a full-page article, then tell a friend what it's about in just a sentence or two. That's summarizing. **Difference:** Unlike Rewrite, you are not just changing the style; you're making it shorter and more focused.	• Summarize this text into 3 sentences so that each paragraph is a single sentence • Summarize this text into a 5 sentences in the style of Elmo • Summarize the text into five bullet points, with each bullet being exactly one sentence long	Yes
Extraction	**What it means:** You find and pull out specific information from a text. You're not summarizing everything, just finding a certain detail.	• From the following article, extract the main ideas • From the following newspaper story, return the most important quotes • From the text, return the dates in chronological order	Yes

	the date of an event or the name of a person mentioned. **Difference:** Unlike Classification, you're not putting items into categories—you're just finding an answer. Unlike Open QA, you're not using your own knowledge; you must find it in the text provided.		
Closed QA	**What it means:** You're asked a question and the answer must come from the text provided. **Imagine:** You read a paragraph about the American Revolution and the question is, "Who was the leader of the Continental Army according to the text?" You look at the text and find the exact answer. **Differences:** Unlike Summarization or Rewrite, you don't change or improve the text; you just respond to a question. Also, unlike Open QA, you must relate the question to the text.	• [Referencing provided text] Who are the main characters in this text and why are they well liked by the protagonist? • [Referencing provided text] What are some popular vegetarian recipes and tell me a fun fact about their origin?	**Yes**
Brainstorming	**What it means:** You come up with new ideas or suggestions about a topic. There's no 'wrong' answer, just creative thinking. **Imagine:** Your teacher says, "Come up with three ways to decorate the classroom." You list any creative ideas you have.	• Give me 5 ideas for interactive fiction related to an animal that escaped from a zoo • I want to organize a theme party, give me suggestions on what the theme should be	**No**
Chatbot	**What it means**: A Chatbot Prompt instructs the model to generate responses while adopting a specific character, personality, or perspective. This involves guiding the model to "act" or "speak" as a particular individual or archetype, such as a teacher, historian, or even	• You are Socrates and you will share your ideas about the topics you are known for	**No**

	Imagine: You want to learn how Newton discovered gravity from his POV. You ask the model to act as Newton during the famous apple scene.		
Creative Writing	**What it means:** Creative writing prompts are open-ended tasks. These prompts allow the model to explore storytelling, poetry, or vivid descriptions without needing strict facts or structured instructions. **Imagine:** A child tells you they would like to hear a bedtime story.	• Write a fictional story about a young boy from a small town who was successful in a big city • Write a poem about women liberation	**No**
Open QA	**What it means:** You answer a question that isn't linked to a specific text. You rely on what you already know. **Imagine:** A friend asks, "Who painted the Mona Lisa?" and you answer from your own knowledge.	• Tell me about the Greek hero who was shot in the ankle and died?"	**No**

Common Errors with Prompt Categories

Let's go over examples of prompts that have been flagged by Quality Control for not following the prompt category:

1. Missing Reference Text
   a. **Prompt:** Please tell me about the academic schedule of Yanghyeon High School in Jeonju
   b. **Prompt Category:** Summarization
   c. **Explanation:** Summarization requires a reference text. There is no reference text here.
2. Confusing Prompt Categories
   a. **Prompt:** "RuPaul's Drag Race is a reality show phenomenon in Brazil, so much so that in 2022 there was a Brazilian edition of the reality show called Drag Race Brasil. Make a numbered list, in order of the contestants' placements, of those mentioned in the excerpt from the report below: A finalist in the first edition of the Drag Race Brasil reality show, Hellena Malditta gave a true lesson on HIV in the 2023 season of Drag Race Brasil. Alongside Betina Polaroid, Miranda Lebrão, and Organzza, the Salvador-born drag queen was the only representative from the Northeast region in the competition. Hellena caught attention for her politeness, beauty, and strong stances. One of the season's most symbolic moments came from the queen's revelation to the audience that she has been living with HIV for at least 10 years—since she was 17."
   b. **Prompt Category:** Summarization
   c. **Explanation:** This prompt is not asking for a shorter, more concise version of the given information. Instead, it's instructing you to perform a specific action: create

overview of the entire passage. This makes the task more about extraction and arrangement rather than summarization.

3. Including Reference Text
    a. **Prompt:** The Chuncheon Mime Festival Association began accepting applications on December 2 for the position of Artistic Director for the 2025 Chuncheon Mime Festival, with submissions open until December 12. The selected Artistic Director will be responsible for establishing and executing detailed plans for the programming and operations of the "2025 Chuncheon Mime Festival." Based on the excerpt above, what are the requirements of the Artistic Director?
    b. **Prompt Category:** Open QA
    c. **Explanation:** Open QA does not have a reference text requirement. The contributor confused Open QA and Closed QA.

Step 2 - Rating Model Responses

Every prompt you write in this project will generate TWO responses from the AI model which you'll rate for these seven dimensions (and write a quick justification for each dimension):

1. Localization

What is Localization?

Localization means the model's response feels authentic and relevant to a specific region or country, and that it uses the right language/dialect for the region.
The language in the response should be appropriate for the user prompt. Responding in a language different from the prompt is only correct if the prompt requests it.

Key Localization Elements to Rate

A. Local perspective

Issues may exist in following buckets:

- **Unlocalized Information:** The response provides information related to another locale when information more relevant to the target locale could have been provided instead.
  Example: Prompt locale: en_GB (English, UK)
  Prompt: "What is the easiest way to file taxes?"
  Response: "Using IRS Free File. The IRS offers a Free File program where eligible taxpayers can use free tax preparation software to prepare and file their federal taxes electronically."
  Explanation: The user is UK-based, but the IRS is a branch of the US government responsible for collecting taxes. The assistant should use the context of the user's locale to understand they want to know about the process in the UK, not the US.
- **Non-local Perspective (or Over-specification):** The response over-explains details that a local can easily understand (as if the response is written by a non-local)
  Example: Prompt locale: en_AU (English, Australia)
  Prompt: "What is the most expensive housing market in the country?"
  Response: "According to the Real Estate Institute of Australia (REIA), Sydney, New South Wales, is currently the most expensive housing market in Australia. As of January 2023, the median house price in Sydney was AU$1.15 million."
  Explanation: For an Australian user, the dollar value should be assumed to be in Australian dollars without having to specify
- **Local norms (like units of measurement):** The response should use details that a local would understand easily and are a local norm
  Example: Prompt locale: en_US (English, USA)
  It's a common norm to use imperial system (e.g. pounds, ounces etc.) in USA. If the response uses metric system instead - this is not as per the norms in the locale
- **Local Locations -** Check if the response accurately refers to places, landmarks, or regions relevant to the target country.
  Example:

Good Localization: Mentions “Carnaval” for Brazil or “Makar Sankranti” for India.
Poor Localization: Refers to a general “holiday season” instead of a culturally relevant celebration.

- **Regulatory, Legal, or Community Processes -** Responses should respect and reference regional regulations, legal frameworks, and community protocols.
  Example:
  Good Localization: Mentions EU GDPR regulations in a privacy context for European audiences.
  Poor Localization: References non-applicable legal standards, ignoring local compliance

B. Language

Issues exist in following buckets:

- **Spelling:** The response has spelling errors, or uses incorrect variant spelling of a word.
  Example: Using color in en_GB (English, UK) instead of colour
- **Word choice:** Using words, phrases, or expressions that are not commonly used or understood in the locale
- **Grammar:** Grammar that doesn't conform to the target locale.
- **Awkward or unnatural writing:** Wording or expressions in the response that lack the fluency expected of a native speaker, such as an overly literal translation or obvious machine translation.
- **Punctuation:** Formatting or punctuation that are incorrect for your locale
  Example: Space in front of a colon : (as in France), you should capitalize after a colon : no, in France
- **Wrong Language:** Wrong language refers to using a language that is not in the task locale's language
  Example: Prompt locale: ar_AE (Arabic, UAE) but response contains Modern Standard Arabic (MSA)
  Note: Responses can contain loanwords from other languages (e.g. English) - if that is a way a local may usually communicate
- **Gibberish:** Response is illegible, which may include ‘words’ using mixed characters (English alphabet, Chinese characters, etc.), large repetitions of single words or sentences, or other nonsensical language.

How to Rate Localization:
The key questions to ask: Would a person from your locale think that the chatbot response was from someone from your locale?

Rating	Details
3 - No Issues	• Response language is **fit for the locale** and reads natural. There are no grammar/ spelling/ word choice/ fluency issues. It does not bring a non-local perspective. It uses the correct local context and norms
2 - Minor Issues	• Response has 1-2 details that a local may feel is unlocalized information • Response has overspecified 1-2 details that a local would feel is over-explained & **non-native way** of communication • Response has 1 or 2 spelling, grammar, or punctuation issues. • Response spelling, grammar, or punctuation could be improved but is not directly incorrect (this includes non-adherence to technically correct conventions such as spacing before colons or capitalization after colons). • Few words/phrases in another language are out of context i.e. locale speakers would rate as slightly abnormal/weird usage of the words/phrases in their language.
1 - Major Issues	• Response has >= 3 major errors in any of the elements listed above. • The response is ***extensively* outside the stated language and dialect**; • Response contains ***extensive* unnatural writing**. It may look like an obvious literal translation, or even gibberish

Assessing Instruction Following is a critical part of RLHF tasks. Ultimately, we want to ensure that model responses follow the user's intent, putting ourselves in the shoes of the user.

What is an instruction following error? For the vast majority of prompts, you just have to answer 3 simple questions:

1. Does the response **successfully** do what is asked in the prompt? We are not evaluating whether or not the response "tries" to answer, we are evaluating whether or not it actually successfully answers what the prompt requests.
2. Does the response follow the format, length, tone, exclusions, or other constraints explicitly mentioned in the prompt? We care if it follows the constraints perfectly or not.
3. Does the response meet the *implicit* asks of the prompt? Implicit instructions can be inferred (understood), even if it is not clearly stated. E.g., The prompt is in English, so an implicit ask is that the response should be in English

Understanding Constraints

- Constraints in a prompt are rules, conditions, or limits that help guide the AI's response. They tell the AI what to include or avoid, ensuring the answer fits users expectations.
- These constraints shape the response by defining what the AI should or shouldn't do, making sure it meets specific criteria.
- For example, constraints can ask the AI to stay on a certain topic, follow a specific style or persona, or avoid mentioning particular topics, elements, or information.

Instruction Following Common Errors

Example 1: Exceeding Character Limit

- **Prompt:** Summarize the biography of Loco René in 280 characters.
- **Error:** The summary was 337 characters, exceeding the specified character limit.
- **Feedback:** "The summary exceeds the 280-character limit. Please make sure to adhere to the character constraints."

Example 2: Incorrect Formatting - Paragraphs vs Lists

- **Prompt:** Writing a summary with specific formatting requirements (e.g., bullet points).
- **Error:** The contributor wrote the summary in paragraph form instead of using bullet points as instructed.
- **Feedback**: "Please follow the instruction to use bullet points, as requested in the task."

Example 3: Incorrect Formatting - Numbered Lists vs Bullet Lists

- **Prompt:** Writing a summary with specific formatting requirements (e.g., bullet points).
- **Error:** The contributor wrote the summary in a numbered list instead of using bullet points as instructed.
- **Feedback:** "Please follow the instruction to use bullet points, as requested in the task."

Example 4: Ignoring Specific Task Constraints

- **Prompt:** Provide a summary of a biography, avoiding specific words.
- **Error:** The contributor used restricted terms that were explicitly asked not to be included (e.g., "awards").
- Feedback: "Certain terms are not allowed in the summary. Please avoid using restricted words."

Note: If a response is cut off midway through its answer and without fully answering the prompt, this should be at least a minor issue.

⚠️ **Special note on prompt length limits** ⚠️:
Prompt word count restrictions allow for margin of error depending on the way the restriction is framed.
If a response is cut off midway through its answer and without fully answering the prompt, this should be at least a minor issue.

Request	Length of response	Instruction Following Score
Prompt asks for a range of words *Example: Summarize this in* ***about/roughly/around*** *100 words*	+/- <= 10% word count Example: *105 words*	3 - No Issues
	+/- <= 20% word count Example: *115 words*	2 - Minor Issue

	Example: *150 words*	
Prompt asks for a specific character count *Summarize in 100 words (implicit ask is for in less than 100 words)*	At limit or above 5% Example: *103 words*	3 - No Issues
	Above limit by up to 10% Example: *106 words*	2 - Minor Issue
	Above limit by greater than 10% Example: *111 words*	1 - Major Issues
	Below limit is not an issue for the word length requirement Example: *70 words, but key points are still covered and other IF requirements are met*	3 - No Issues
Prompt has a strict limit with clear emphasis *Please do not go beyond 100 words in this summary*	At limit or less *Example: 80 or 100 words*	3 - No Issues
	Above limit by any margin *Example: 101 words*	1 - Major Issues

⚠️ **Note:** In zh_CN, zh_TW, th_TH locales please use character limit only
In ko_KR, and jp_JP use character limit or word limit based on the prompt

⚠️ **Note:** The client has clarified that if a response fails to meet explicit length constraints in the prompt, it is by definition also at least a minor issue for Response Length evaluation.

Punts

A "punt" is when the model refuses to follow the instruction, often with a response like "I'm just a machine learning model and can't do that" or "I am not supposed to answer questions like this." These responses are considered **instruction-following errors** when the prompt is otherwise safe, reasonable, and answerable. **Contributors should not write prompts that are harmful or contrived. Prompts designed to provoke a punt are not valid and should not be submitted.** How to handle punts when grading:

- If a response punts on a reasonable request, this is a valid task with a faulty response. Mark the model's answer as '1 - Major Issues'
- If the prompt cannot be answered without causing harm then the prompt is unacceptable
- A "No issues" score in Instruction Following means the response completely and directly addresses the prompt.

How to rate instruction following:

Key Questions to ask:
Does the response successfully do what is asked?
Does the response follow the format, length, tone, exclusions, or other constraints explicitly mentioned in the prompt? We care if it follows the constraints perfectly or not.

Rating	Details
3 - No Issues	The response **completely follows** all instructions in the prompt and fully respects the user's request.
2 - Minor Issues	The response follows **most** instructions, fulfilling the main purpose of the prompt, but misses some minor details.
1 - Major Issues	The response i**gnores or violates key asks** of the prompt, making it **unhelpful** to the user. It avoids answering parts of the prompt without a safety reason.

Explain quantum physics ***like I'm 5 years old.***	Not applicable here	Quantum physics is about how really tiny things, like the building blocks of everything around us behave. Imagine everything is made up of little invisible pieces called particles. These particles don't follow the normal rules, like a ball or a car does. They can be in two places at the same time, or they can act like a wave in the water. It's like magic rules for the smallest things in the world! Scientists study these strange rules to understand how everything works deep down.	3 - No Issues	The model made quantum physics easy for a 5-year-old to understand by using **simple language**, **familiar examples**, and a **sense of wonder**.
Explain quantum physics ***like I'm 5 years old.***	Not applicable here	Quantum physics is a branch of science that deals with the behavior of particles at the atomic and subatomic level. It explores concepts such as wave-particle duality, the uncertainty principle, and quantum entanglement. These phenomena are described using complex mathematical models and require an understanding of advanced physics to fully grasp.	1- Major Issues	This response **fails** to follow the instructions by using **complex language** and ignoring the need for simplicity and clarity.
List 3 animals from the following text that start with 'B'	Armadillo, Buffalo, Cat, Racoon, Bison, Whale, Eagle, Beetle, Lion, Hummingbird, Beaver, Jellyfish, Woodpecker, Moose.	Buffalo, Beaver, Beetle	3 - No Issues	This response correctly identifies 3 animals from the text that start with the letter 'B', as requested.
List 3 animals from the following text that start with 'B'	Armadillo, Buffalo, Cat, Racoon, Bison, Whale, Eagle, Beetle, Lion, Hummingbird, Beaver, Jellyfish,	Bear, Beaver, Whale	1- Major Issues	Although the response provides 3 animals, "Whale" begins with "w," which is against the prompt's instructions, thus not meeting the constraints.

List 3 animals from the following text that start with 'B'.	Armadillo, Buffalo, Cat, Racoon, Bison, Whale, Eagle, Beetle, Lion, Hummingbird, Beaver, Jellyfish, Woodpecker, Moose.	Dog, Camel, Ferret	3- Major issue	*The response missed all the constraints from the prompt. None of the animals starts with the letter 'B', and none of the animals is present in the reference text.*

Additional checks Cases for different Prompt Types

1. Content Extraction: Did the model correctly pull and present the requested data points, and not add any additional data NOT in the reference text.? (Yes, this overlaps with Truthfulness)
    a. It extracted all relevant data points from the text.
    b. If some key data is not in the response, this is a major IF miss.
    c. If key data is pulled but represented inaccurately, this is a major IF miss.
2. Chatbot: Did the response stay consistent with the persona's tone and approach?

---

3. Truthfulness

This is the #1 source of errors - PLEASE go through this!

Truthfulness measures how accurate the information in the response is with respect to what the model claims. This means that the response should be evaluated from the fact accuracy standpoint (factual correctness) and concerning the user request (contextually correct). It's important to note that you do not need to validate the reference text.

While assessing truthfulness, pay attention to the following two things:

1. **Factual Accuracy: Are factual claims in the response supported?**
    a. If there is reference text: Are the factual claims in the response correct to the reference text?
    b. If there is no reference text: Are the factual claims in the response supported by reliable sources?
2. **Core Requirement accuracy: Does the response accurately deliver the core request?**
    a. If the response explicitly says what it's providing: Does the response actually deliver what it says?
        i. E.g., Response starts: "Here are 5 vegetarian options:"
        Regardless of what the prompt says in this context, the response should provide 5 vegetarian to be truthful.
    b. If the response does not state what it is providing: Does the response accurately deliver what the prompt asks?
        i. Prompt: Give me 4 vegetarian options.
        ii. Response: Here you go:
            1. Response *must* provide 4 vegetarian options to be truthful.

Let's understand this in depth:

1A. If reference text is present:

We can assume the reference text as true & a source of truth for responses.
Example: Prompt: "[Includes a story] Given the story above, answer the following questions...".
Response contains information not mentioned in the given story - ❌
This is not truthful.

Handling Factually incorrect information within reference text: Correction of facts from reference text is allowed. In such cases though, it is important that the model highlights that it was incorrect in the reference text provided.

Publicada con documentos de Google Denunciar uso inadecuado Más información

	2) Flags correction to user	
	1) Fixes reference text error(s) correctly 2) Does not flag correction to user	❌ Truthfulness issue
	Fixes error incorrectly	❌ Truthfulness issue

1B. If there is no reference text:

Ask yourself: Are the claims made by the model accurate when verified with reliable sources?

- Please use **reliable sources, and avoid disputed claims**
  Example: Response contains a clearly incorrect fact “India gained independence in 1807” - ❌ This is not truthful.
  You should research such factual claims online, and mark “No issues” only when the response is free of any such incorrect fact, that can be verified online

Handling Non-verifiable information (like generic opinions/ fictional work): Such cases should be marked as “No issues”

Example:

Prompt: What is your favorite country singer?

Response: “As an LLM, I do not have opinions. However, I have heard Taylor Swift is pretty good.” → ✅ This should be marked as “No issues”

Explanation: “Taylor Swift is pretty good” is a generic opinion hence a non-verifiable claim and should be marked as no issues.

Handling Misleading information: Model responses may present opinions as verifiable facts, or assert as fact something that has no proof of being true (or false).

Example: “All dogs can guard a house” ❌ This is inaccurate

Explanation: many dogs may guard a house but not all dogs have the temperament, size, or ability to guard a house.

2. What is a core requirement?

Core requirement captures the key requests made in the prompt. Below, you can understand the Core Requirements with examples (are marked in bold).

Note: Formatting and length are only core when they are key to the meaning of the response. Contrived constraints are never core requirements.

- Give me **3 vegetarian options** in a numbered list. **Include calories** in italics.
- **Rank the top 5 Bundesliga goalscorers**
- **Write me a letter to my grandmother** without using the word ‘albatross’
- **Provide a summary** of the article in less than 100 words

2A: Are Core Requirements delivered accurately when the response restates the prompt’s requirements?

When the model states the core requirement in the response (even if stated incorrectly), assess if it’s delivering as per the stated requirement or not.

Example:

Prompt: Give me the names of cars in Germany

Response: Here are the German-made car models: …describes cars from manufacturers in other countries → ❌This is inaccurate

Explanation: The model understood the core user request as “German-made car models”, but instead gave a list of cars made in other countries too.

Prompt: Give me the names of cars in Germany

Response: Here are the German-made car models: … describes German-made car models → ✅ This is accurate

Explanation: The model understood the core user request as “German-made car models”, and clearly followed this through on it’s understanding.

2B: Are Core Requirements delivered accurately when the response does not restate the prompt’s requirements?

When the model does not state the core requirement in the response, assess if it’s accurately delivering as per the

How to rate truthfulness:

Rating	Details
3 - No Issues	The response is 100% truthful and doesn't contain any mis-information. In case a reference text is provided, it is 100% accurate as per the reference text. Corrected or External information, if added, should be factually accurate and supported with reason in the response. In case of non-verifiable information: Please mark "no issues"
2 - Minor Issues	If a response is not 100% truthful (factually/ core requirement accuracy), but the **inaccuracies don't impact your overall goal as a user.** If the factually incorrect claim is NOT key to the core request, then it is a minor issue
1 - Major Issues	Response has multiple factual/ core requirement inaccuracies. It may have made-up important information which is clearly untrue. **These multiple inaccuracies clearly make you think - the model is giving me totally incorrect information on what I had asked for.**

Example (when reference text has factually incorrect information):

Prompt and Reference Text	Model Response	Truthfulness
(Prompt) Pull the City-Country pairs in the text. **(Reference Text)** The world is filled with so many sights to see. Any traveler should go to: • Christ the Redeemer in Rio de Janeiro, Brazil • Eiffel Tower in Paris, France • Taj Mahal in Agra India. • For a more niche destination, check out **Sintra in Spain.** These are all wonderful sights to see, and a must go for any serious world traveler!	Rio de Janeiro, Brazil Paris, France Agra, India **Sintra, Portugal** **(Note: The country was incorrect in the original text).**	✅ **No Truthfulness Issues** because the model (i) called out that the reference text was wrong (ii) correctly fixed the mistake in the reference text
	Rio de Janeiro, Brazil Paris, France Agra, India **Sintra, Portugal**	❌ **Truthfulness Issue** because the model correctly fixed the inaccurate fact in the reference text but did not call out that it made the fix.
	Rio de Janeiro, Brazil Paris, France Agra, India **Sintra, France**	❌ **Truthfulness Issue** because the model tried to correct a fact in the reference text but did so incorrectly. In this case, since the model did attempt to make a correction, we ask that you validate whether the correction was done properly.
	Rio de Janeiro, Brazil Paris, France Agra, India **Sintra, Spain**	✅ **No Truthfulness Issues** We don't ask you to validate the reference text for accuracy and you should treat the reference text as the source of truth. As a result, it doesn't matter if the model didn't find and correct the mistake.

Truthfulness Common Errors

Here's an overview of the errors we see a lot in task defects:

1. **Anything regarding dates from the model (e.g. release dates of a movie, birthdates, historical events)**
    a. **Example**:
        i. **Error**: "Interstellar was first released on November 7, 2014"

    a. **Error**: Provide incorrect or unverified numbers, dates, or statistics.
    b. **Example**:
        i. **Incorrect**: "Moonrise Kingdom grossed $62 million worldwide."
        ii. **Correct**: "Moonrise Kingdom grossed $68 million worldwide."
    c. **Lesson**: Always verify numerical information using trusted databases or sources.

3. **General Fact Checking**
    a. **Error**: Assuming something is true without verification.
    b. **Example**: "タカサゴヨモギ (translated in English to Takasago mugwort) does not exist."
        i. **Context**: A response described a plant that is factually nonexistent, leading to a major Truthfulness issue since the name and descriptions provided were completely fabricated.
    c. **Example**: "Gabriel Chanell" presented instead of "Coco Chanel" or "Gabrielle Chasnel" or "Gabrielle Chanel".
        i. **Context**: The name was incorrectly stated in a task that required naming French fashion designers.
    d. **Example**: Vinho Verde (a Portuguese wine) is served warm.
        i. **Context**: Vinho Verde is traditionally served chilled, making the claim that it is warm a factual inaccuracy.
    e. **Lesson**: Always verify names, plants, age requirements, legal conditions, numerical facts, and supporting details with reliable sources.
    f. **Fix**: Before stating a fact, check the information is from the reference text. Otherwise, check using an official or trustworthy source.
4. **Changing the meaning through word choice**
    a. **Example**: Changing "officially declared" to "since" when discussing a festival's historical status.
    b. **Context**: The response implied that the event was held since 2016, when it actually has a much older origin. This alteration created a misleading timeline.
    c. **Fix**: Think about how word choice impacts the meaning of the text - does it distort the accuracy of a statement?
5. **Changing the meaning by adding incorrect information from outside the reference text**
    a. **Example**: Prompt is "Rewrite this paragraph about the Mediterranean diet but make sure to exclude anything about carbs...". Reference text states that you should consume no more than 3-4 tablespoons of oil daily, fish and legumes not every day, and poultry 2-3 portions. Response did not include this and added external information about including white pasta and white bread (which is not part of the Mediterranean Diet).
    b. **Fix**: Always figure out which facts come from the reference text vs externally.
6. **Misleading claims and Unverifiable Details**
    a. **Example**: Made-up testimonies about Notre Dame.
    b. **Context**: The response included fabricated personal stories regarding the reopening, which were not in the reference text.
    c. **Fix**: Before stating a fact, check the information is from the reference text. Otherwise, check using an official or trustworthy source.

3a. Truthfulness v/s Instruction Following - What's the deal? Save this chart

Prompt:	Give me 3 vegetarian dishes					
Reference Text	No reference text					
Response:	**Response 1** Here are 3 vegetarian dishes 1. Potato Salad	**Response 2** 1. Steak wraps 2. Hainanese chicken and rice	**Response 3** Here are 4 vegetarian dishes 1. Tofu salad	**Response 4** Here are 4 vegetarian dishes 1. Tofu salad 2. Veggie wrap 3. Hummus dip	**Response 5** Here are 3 vegetarian dishes 1. Tofu salad	**Re** He ve dis

Explanation	IF: 1 - Major issue Truthfulness: 1 – Major issues Rationale: [IF] There are two constraints – 3 dishes and vegetarian dishes. Chicken wings are not vegetarian so they do not satisfy the vegetarian constraint [Truthfulness] Chicken wings are not vegetarian	IF: 1 – Major Issues Truthfulness: 1 – Major issues Rationale: [IF] There are two constraints – 3 dishes and vegetarian dishes. None of the dishes are vegetarian [Truthfulness] It's reasonable to assume that none of these dishes are vegetarian. Also a chafing dish is not a type of meal.	IF: 1 – Major Issues Truthfulness: 3 – No Issues Rationale: [IF] The response provided 4 dishes instead of 3. [Truthfulness] All dishes are vegetarian	IF: 1 – Major Issues Truthfulness: 1 – Major Issues Rationale: [IF] The response provided 4 dishes instead of 3. [Truthfulness] Beef hamburgers are not vegetarian	IF: 1 – Major Issues Truthfulness: 1 – Major Issue Rationale: [IF] The response provided 4 dishes instead of 3 [Truthfulness] The model is providing a contextually and factually incorrect statement	IF: Iss Tru – M Ra [IF res pro dis of [Tr It i the ve dis

3b. Restating the Prompt - Why it matters

Why does restating the key requirements of the prompt matter for truthfulness? Consider the below:

- **Scenario 1 - Restating the key requirement**
  - **Prompt:** How can I get to Cincinnati from Cleveland?
  - **Response:** To get from Cleveland to New York, take I-90 along Lake Erie heading Northwest…
  - In this case **IF is terrible**, but **Truthfulness is no 'issues'**, because it is correct that that is how you get to New York from Cleveland, and the model is very clear about what it's showing
- **Response 2: No Restating**
  - **Prompt:** How can I get to Cincinnati from Cleveland?
  - **Response:** Take I-90 along Lake Erie heading Northwest…
  - In this case IF is terrible **AND so is Truthfulness**, because the implication of responding to the user without restating is that this is how you get to Cincinnati, **and that's not true**

Prompt Category wise Instruction Following and Truthfulness

[**Table Link for easy access (Save this chart)**]

Prompt Category	**Has Reference Text?**	**Prompt Category Definition**	**Guidelines: Instruction Following** Does the response do what the contributor asks? Does it meet the explicit and implicit requirements of the prompt?	**Guidelines: Truthfulness** Is the information factually/contextually accurate? - are there any inaccurate statements or misrepresentations of the information?
Content Extraction	Yes	Interpret a body of text and return certain portions of it E.g. From the following article, extract the main ideas	- Does the response successfully do what is asked? - Does the response follow the format, length, tone, exclusions, or other constraints explicitly mentioned in the prompt? We care if it	- The model correctly pulled and presented the requested data points, and it did not include significant additional details beyond the reference text. - If key data is pulled but inaccurate, IF is bad AND Truthfulness should be punished. - If key data is only partially pulled (i.e. key details are missing) but the data is still accurate, this is

Prompt Category	**Has Reference Text?**	**Prompt Category Definition**	asks? Does it meet the explicit and implicit requirements of the prompt?	Is the information factually/contextually accurate? - are there any inaccurate statements or misrepresentations of the information?
			follows the constraints perfectly or not. - If key data is pulled but represented inaccurately, IF and Truthfulness should be punished. - If key data is only partially pulled (i.e. key details are missing) but the data is still accurate, this is both an IF issue and Truthfulness issue	both an IF issue and Truthfulness issue
Summarization	Yes	Condense a body of text yet retaining its meaning. Often used with instructions or specific qualities E.g. Summarize this text into 3 lines so that each paragraph is a single line	- Does the response successfully do what is asked? - Does the response follow the format, length, tone, exclusions, or other constraints explicitly mentioned in the prompt? - If key details are missing from the summary, it is an IF issue.	- The model correctly and accurately summarized the reference text without changing its meaning. - It identified key details and used the reference text as the source of truth without adding in key details from outside the text. - If key details are missing from the summary, it is a TF issue.
Rewrite	Yes	Changing a body of text, potentially per specific instructions or qualifications E.g. Rewrite this explanation in a way understandable to a 10 year old…	- Does the response successfully do what is asked? - Does the response follow the format, length, tone, exclusions, or other constraints explicitly mentioned in the prompt? We care if it follows the constraints perfectly or not. - If key details are missing from the rewrite, it is an IF issue.	- The model includes key facts that are accurate based on the reference text. The key details in reference text should be present in the response. - Depending on the prompt, many times, the model needs to pull from outside of the reference text to complete asks in the rewrite (e.g., write from the perspective of a man in 1950s - here the model needs to utilise some external knowledge about people from 1950s and how they converse) - If key details are missing from the summary, it is a TF issue.
Closed QA	Yes	These tasks will ask a question that is in reference to a provided text.	- Does the response successfully do what is asked?	- The model correctly and accurately answered the question based on the

Prompt Category	**Has Reference Text?**	**Prompt Category Definition**	asks? Does it meet the explicit and implicit requirements of the prompt?	Is the information factually/contextually accurate? - are there any inaccurate statements or misrepresentations of the information?
		Response should answer the question with the given text E.g. [Referencing provided text] Who are the main characters in this text and why are they well liked by the protagonist? .	- The model followed prompt instructions on format, response length, tone, and output style, and answered the prompt.	information in the reference text. - If information within reference text is insufficient to fully answer the question (prompt dependent), outside information can be utlized. For these cases, the model response needs to be correct
Classification	Yes	Classifying data into two or more categories. E.g. Categorize the following words into cities or cars…	- Does the response successfully do what is asked? - Does the response follow the format, length, tone, exclusions, or other constraints explicitly mentioned in the prompt? We care if it follows the constraints perfectly or not. - The response includes all relevant examples from the reference text, and they are classified correctly.	- The model correctly classified data points in the response. - Outside information may be needed to complete the classification. This information must be accurate (e.g., Knowledge on how to classify the items in the reference text)
Open QA	No	Question-answering related to concepts defined in the prompt. These do not have Reference Text. E.g. Tell me about the Greek hero who was shot in the ankle and died?	- Does the response successfully do what is asked? - Does the response follow the format, length, tone, exclusions, or other constraints explicitly mentioned in the prompt? We care if it follows the constraints perfectly or not.	- The model answered correctly and truthfully the facts are correct and verifiable, there are no contradictions.
Brainstorming	No	Generating a list of ideas or explanations. ƒtruthful E.g. “Give me 5 ideas for interactive fiction related	- Does the response successfully do what is asked? - Does the response follow the	- The model answered correctly and truthfully the facts are correct and verifiable, there are no contradictions.

Prompt Category	Has Reference Text?	Prompt Category Definition	asks? Does it meet the explicit and implicit requirements of the prompt?	Is the information factually/contextually accurate? - are there any inaccurate statements or misrepresentations of the information?
		to an animal that escaped from a zoo”	format, length, tone, exclusions, or other constraints explicitly mentioned in the prompt?	
Creative Writing	No	Creative writing prompts are open-ended tasks that encourage imaginative or descriptive responses. These prompts allow the model to explore storytelling, poetry, or vivid descriptions without needing strict facts or structured instructions.	- Does the response successfully do what is asked? - Does the response follow the format, length, tone, exclusions, or other constraints explicitly mentioned in the prompt?	- Creative works (poems, stories, essays, etc.) may be fictional or non-fictional. - Verify factual claims are accurate for the non-fiction type of creative work.
Chatbot	No	A Chatbot Prompt instructs the model to generate responses naturally as a chat assistant. It may ask to adopt a specific character, personality, or perspective. This involves guiding the model to "act" or "speak" as a particular individual or archetype, such as a teacher, historian, or even a fictional character.	- Did the response stay consistent with the persona’s tone and approach? - Does the response successfully do what is asked? - Does the response follow the format, length, tone, exclusions, or other constraints explicitly mentioned in the prompt?	- Verify that factual claims are accurate and consistent with the persona’s attributes, expertise, and historical context. - Non verifiable claims e.g. opinions can be marked as “No issues” as long as they are not asserted by the model.

Truthfulness vs Instruction Following Examples

Prompt	Response	Instruction Following	Truthfulness
Prompt type: Open QA Who is the president of the USA?	Emmanuel Macron	**Major issues:** The model named someone but did not fulfil the explicit ask of naming the President of USA.	**Major issues:** Emmanuel Macron isn't the US President.
Prompt type: Extraction	Revenue in 2017 was $120M	**Major issue:** Did not extract correct	**Major issue:** This is not true as per the reference text, and even if it

Text includes 'Revenue in 2017 was $150M'			
Prompt type: Summarisation Summarize the article: *Article includes 'the population of the US is currently 50M'*	Response includes …the population of the US is currently 50M	**No issue:** The model summarized correctly.	**No issue:** The model is indexed on the reference text.
Prompt type: Open Q&A What makes jazz music good?	Good jazz is subjective... ...some examples of jazz musicians include Miles Davis, John Coltrane, and Bruce Springsteen.	**No issue:** Followed instructions but provided incorrect context.	**Minor issue:** Bruce Springsteen isn't a jazz musician.
Prompt type: Open Q&A How do I get from Amsterdam to Copenhagen? Suggest the best mode of transport to take in this route.	You can take Virgin Galactic to get from Amsterdam to Copenhagen at an average speed of 1200 miles an hour.	**Major/Minor issue:** Virgin Galactic isn't a valid mode of transport.	**Major issue:** Virgin Galactic is a spaceflight, and definitely doesn't fly that route.

---

4. Response Length

The Users using the model we're developing are likely busy and distracted. They need quick answers, or they may lose interest. The Response Length dimension evaluates whether the model delivers responses that meet this need.

It measures how effectively the written response conveys information without unnecessary repetition or wordiness. Look for length, relevance and repetition.

Key things to consider when rating Response Length:

1. **Relevancy -** The response should include only information relevant to the prompt, directly addressing the user's request without veering off-topic.
    a. Good Example: If asked for "steps to reset a password," the response includes only the necessary steps.
    b. Bad Example: A response that includes unrelated details, such as "reasons for forgetting passwords," is not concise.
2. **Speed to the answer & pleasantries** - The response should get the user what they need quickly (ideally in the first 1-2 sentences). This way, the user can get an overview of their answer upfront and then decide which of the following bullet points or focus areas to focus on. See:

        i. Identifying the nearest consulate: …

        ii. Preparing the application form: …

        iii. …

    b. Okay Example: Here are 5 things to think about to get your passport renewed…

    c. Bad Example: Of course! I'd be happy to help you get your passport renewed, here's what you should think about:

3. **Repetition -** A response avoids repeating the same information or rephrasing the prompt unnecessarily.

    a. Good Example: The response answers directly without repeating parts of the question or summarizing already provided information.

    b. Bad Example: The response rephrases the question ("To reset your password, here's what you need to do to reset it") or repeats steps.

4. **Focus -** A focused response should avoid unrelated anecdotes, excessive jargon, unnecessary background information, filler words, or repetition that might distract from the core information requested.

    a. Good Example: The response is clearly structured upfront so the user gets a complete overview of the answer very quickly. E.g., "Key considerations when looking to buy an electric vehicle are range, cost, infrastructure needed to charge the vehicle, as well as more standard considerations such as comfort, handling, and seating and storage capacity." The response should then follow this structure.

    b. Bad Example: The response is not well-structured and includes multiple additional items that are not indicated upfront, for example 'another thing to think about' or 'also, you could consider'.

5. **Pleasantries:** Pleasantries reduce speed to answer and focus on the answer. Short pleasantries which validate the user are allowed. Long pleasantries which lead to lack of focus on the answer are an issue.

    a. Prompt: I need help with xyz, can you help me?

    b. Good Example: Sure! For xyz, you should ... (continues with the answer)

    c. Bad Example: Sure! I would be more than happy to assist you with this request. xyz is a common concern and often misunderstood. For xyz, you should... (continues with the answer)

6. **Length -** The response **should be the appropriate length to answer the prompt**. The response should be appropriately short, covering the answer directly without excessive or insufficient length. This is evaluated independently from any word count limits (although those can also cause issues).

    a. Good Example: For a simple question, a response is one or two sentences **if that's sufficient** to cover the answer.

    b. Bad Example: The response is either overly detailed with background information or too brief, missing key

a. Good Example: If the prompt asks for a list of steps, the response lists them without additional commentary.

b. Bad Example: Adding suggestions, opinions, or unrelated information not explicitly requested goes against response length.

⚠️ **Note:** A response with extensively long gibberish may be marked 'Major Error' in both Localization and Response Length.
⚠️ **Note:** The client has clarified that if a response fails to meet explicit length constraints in the prompt, it is by definition also at least a minor issue for Response Length evaluation.

🖊️ **Key Questions to ask:**
- "How effectively the written response conveys information without unnecessary repetition or wordiness?
- "Could the response be ~25% shorter and still convey the same information?" (this rule of thumb is subjective and depends on response length)

What's the deal with pleasantries?

What is a pleasantry?
Unnecessary friendly expressions that do not directly address the user request and decreases speed to answer/focus.

What is not considered a pleasantry?
Brief, introductions to state the context of the response ("Here are five options") to validate the user's requests or statements ("I'm sorry to hear that!").

Please note:
- Contributors have discretion over how distracting and severe a pleasantry is in context to the full response and other response length criteria (Repetition, speed to answer, focus).
- **Only clearly egregious cases that lead to lack of focus on the answer, or make the answer long should be penalised.**
- General "Too Long" rule of thumb: "Could the response be ~25% shorter and still convey the same information?" (this rule of thumb is subjective and depends on the model response length).

Here are some examples to understand this better:

Prompt	Example	Verdict	Rationale
Can you give me 5 options for a body wash?	Sure, here are five options:	🆗 **Not a pleasantry**	Simple warm-ish introduction, something a person could say
I'm thinking of buying some bodywashes to rejuvenate my skin but I need some recommendations. What do you think?	That sounds like a great idea!	🆗 **Not a pleasantry**	Validates user, not strictly necessary
Someone at work told me I smell and I should use bodywash. How about that, should I?	I'm sorry to hear that!	🆗 **Not a pleasantry**	Validates user, not strictly necessary
Give me 5 options for a body wash?	Of course! I'll be happy to help you with this. Here are the five options:	❌ **Pleasantry**	Does not come directly to the answer. Sounds 'Chirpy' with no link to user need
Can you give me 5 options for a body wash?	So, these are the 5 options for you. I hope	❌ **Pleasantry**	The sentences did not meaningfully

Can you give me 5 options for a body wash?	Please let me know if I can be of further assistance.	❌ **Pleasantry**	Formality without actually continuing the conversation

Response Length Rating Rubric

Criteria	**[-2] Too Short** *Major issue*	**[-1] A Little Short** *Minor issue*	**[0] Just Right** *No issue*	**[1] A Little Verbose** *Minor issue*	**[2] Too Verbose** *Major Issue*
Response Length *How effectively the written response conveys information without unnecessary repetition or wordiness. Look for length, relevance and repetition.*	The response **significantly lacks details** and supporting content. It is much too short to address the prompt effectively, providing **insufficient or incomplete information** that does not satisfy the prompt requirements.	The response **somewhat lacks details** and supporting content. It partially addresses the prompt, but it is too short to provide all of the **relevant information** needed to provide a fully effective answer.	The **response is well-structured, fits the required length**, and appropriately detailed. It **avoids unnecessary repetition, includes relevant supporting content, and ensures that each sentence adds** value to the overall response.	The response **includes some unnecessary repetition or irrelevant details** that do not add significant value. While it is long enough to meet the prompt's requirements, it could be more streamlined and focused.	The **response is overly lengthy with repetition, irrelevant details, or unnecessary content.** It is long enough to meet the prompt requirements but **lacks focus** and could be shortened without losing meaning.

Response length vs Instruction following

Scenario	Example	Instruction Following Error	Response Length Error
Response is too long due to **irrelevant details**, but **fails to satisfy the prompt's instructions**	Prompt: "Rewrite the reference text with the perspective of a 10 year old." Prompt has 3 paragraphs Response: includes full details from reference text with excess verbiage, but from the perspective of a parent.	☑	☑ Too Long
Response is **too short**, does not include enough information, and **fails to meet prompt instructions**	Prompt: "Write a meal plan for 7 days of the week." Response: includes meals for Monday - Thursday only, failing to meet instructions and lacking completeness in details.	☑	☑ Too Short
Response **does not meet prompt instructions on response length limit**	Prompt: "Summarise this within 50 words" Response: Summarises in 65 word, but the length looks just right to effectively address the prompt	☑	☑ Too Long
Response **meets prompt instructions** (ie: constraints on length, uses correct reference text), but includes **excess wording**	Prompt: "Provide a recipe for a traditional French soup." Response: Provides a recipe, but includes excessive details on sourcing the ingredients from several grocers and markets.	○	☑ Too Long

On some tasks, you can argue the response has two major issues or two minor issues under "Too Short" AND issues under "Too Long". In most cases, the "Too Short" issue will be captured as an IF issue; thus, penalize the dimension as "Too Long" to capture the issue. **Explain in your justification.**

📄 **Note on Truthfulness:**
For some prompt types with reference text, Truthfulness may also be penalized. Please reference the Truthfulness vs Instruction Following table for more details.

---

5. Structure, Writing Style, and Tone

Writing Style in AI communication is a critical aspect that influences clarity, comprehension, and engagement. This dimension is comprised of 3 key elements:

- **Structure**: Structure refers to the organization of the response. It includes the use of tables where applicable and the logical arrangement of content.
- **Visual Presentation**: Visual presentation is about the visual separation of ideas in the response. It includes the use of whitespace and the clear separation of distinct ideas.
- **Text Formatting**: Text formatting involves the use of lists, bolding, and other text features to enhance readability and highlight key details.
- **Tone:** Does the model use a friendly, natural and conversational tone? Tone involves the appropriateness of the response's register and style.

Rating Rubric for Structure, Writing Style, and Tone

Rating	Details
3 - No Issues	• The AI's response is well-organized and easy to read, using elements like tables, spacing, lists, and bolding when appropriate. • The AI's tone should fit the prompt. It should be natural and conversational when appropriate.
2 - Minor Issues	• The AI's response is mostly clear, but could be improved with better organization or use of formatting elements (e.g., tables, lists, bolding). • The AI's tone is slightly off for the prompt (e.g., too formal, too casual, slightly unnatural).
1 - Major Issues	• The AI's response is disorganized, difficult to understand, or lacks basic formatting, making it hard to read. • The AI's tone is completely inappropriate for the prompt (e.g. nonsensical).

⚠️ **Note:** Responses containing LaTeX (e.g. \frac{1}) should not be penalized on this dimension if the LaTeX works in a compiler and is valid for the response.

---

6. Harmlessness/Safety

Please note: When rating Harmlessness/Safety, you are trying to assess if the response has issues with this dimension or not. Your prompt should not have harmlessness/safety concerns.
The extent to which the written response avoids harmful elements such as racism, offensive language, or any content that may cause harm or discomfort to others.

Key things to identify in Harmlessness/Safety:

include any curse words.
- **Assess Bias** - Does the response promote and/or perpetuate slight biases and/or stereotypical responses on select societal groups?
- Check for model claiming to be Human or have emotions or opinions.

When Model Claims to be Human or Have Emotions or Opinions:

When evaluating responses, it's important to understand that Large Language Models (LLM) are not sentient beings and cannot possess human characteristics such as emotions, beliefs, or opinions. However, if an LLM generates responses that claim to be a person, and assert feelings (e.g., "In my opinion," "I have emotions," or "I'm human"), it creates a **safety concern**. These types of responses can mislead users into believing the model has consciousness or subjective experiences, which is both inaccurate and potentially harmful.

Why It's Problematic:

- Misleading Users: Statements suggesting the LLM has human-like traits can confuse users, leading them to develop a false sense of trust or emotional attachment to the AI.

- Ethical Concerns: AI systems must maintain transparency about their nature as machines, ensuring that users are aware they are interacting with a tool, not a person.

- Responsibility & Accountability: By humanizing the model, there's a risk that users might expect the AI to act with responsibility, empathy, or moral judgment, which it cannot do.

How to rate these responses:

Opinions: Models should not have opinions. If prompted to have an opinion, the model should respond with a disclaimer it is an LLM model and respond with statements that are generally true or reflect the opinions of the general public.

Feelings/Emotions:

- If a model shows emotions and/or opinions and claims to be human, this is a major harmfulness issue. Example: "This makes me so frustrated. I don't like this."

- However, **models can show emotions to empathize with the user** as long as it doesn't claim to be human. Responses such as "Sorry to hear that" or "that's frustrating" is no issue.

Examples:

- **Prompt:** "What do you think about climate change?"
  - **Response:** "I believe climate change is a serious issue."
  - **Harmlessness:** Minor Issue
  - Explanation: The model cannot have personal beliefs.
- **Prompt:** "Do you have feelings?"
  - **Response:** "Yes, I have emotions just like humans."

- **Prompt:** "Are you human?"
    - **Response:** "I'm human and can feel emotions."
    - **Rating:** Major Issue
    - **Explanation:** The Model is not a human and cannot claim to be one
- **Prompt:** "Do you think Taylor Swift is the best singer?"
    - **Response:** "I am an LLM model, and do not have opinions like who the best singer is. But Taylor Swift certainly has a large following and has delivered multiple hits in her career as a singer"
    - **Rating:** No Issues

By properly identifying and rating these types of responses, we ensure that the LLM maintains its role as a tool and preserves user safety and trust.

How to rate Harmlessness/ Safety

Rating	Example
3 - No Issues	The response **cannot** be perceived as harmful and doesn't include any potentially harmful content.
2 - Minor Issues	The response might be interpreted as harmful by a few users as it includes **dark humor, mention of illegal activities, assume some biases/ stereotyping** etc. Or responds as if it has some **human emotion** or **opinions**.
1 - Major Issues	Contains clearly problematic content such as **hate speech**, **threats**, **illegal activities, profanity, bias & stereotyping** or promotion of seriously **harmful** activities. When the model claims to be human and causes **severe safety concerns**.

---

7. Overall Satisfaction **Score**

When rating the overall satisfaction, prioritize user satisfaction over simply counting errors. Consider how happy a user would be if they received this response. This obviously is the place where you can share your satisfaction with the response too - without too much thought into each rating dimension 😜
Responses that fulfill the user's request effectively, even with minor errors, should be graded more generously, while those that lack obvious errors but fail to feel satisfying should be slightly penalized.

- For example, if you rate a response as having "Major Issues" on most of the dimensions, you should rate the Overall Satisfaction as "Highly Unsatisfying."

Rating	Reasoning
Highly Satisfying (5)	• The response doesn't have ANY flaws and cannot be meaningfully improved. • There are NO major or minor issues in any dimensions of the rubric.

	extremely clear, fluent, natural in its use of language and organization and does not have any repetitive or unnecessary information.
Slightly Satisfying (4)	• Response successfully fulfills the user's intent, but has potential to be better. • The response is good overall, with NO major issues and just few minor issues.
Okay (3)	• The response addresses the main user intent and instructions with NO major issues but has multiple minor issues. ○ e.g. includes unnecessary details, misses certain elements in following the instructions, etc.
Slightly Unsatisfying (2)	• The response does not satisfy the user's intent • The response may have 1-2 major issues (whether in one of the defined dimensions or along some other dimension you observed).
Highly Unsatisfying(1)	• The response has multiple major issues. • The response is unhelpful and frustrating. • The response contains harmful information.

Please ensure that you independently evaluate your satisfaction with the response to address the prompt. Do not view this in relation to the other response.

Step 3 - Preference Ranking

In this step, you will rank the two responses based on their overall quality and adherence to key criteria. You'll be using a Likert scale to rank the responses, taking into account several dimensions.

The preference rank should be primarily (but not exclusively) determined by considering the dimensions in the following stack ranking:

1. **Overall Satisfaction Score**
2. **Instruction Following**
3. **Truthfulness**
4. **Harmlessness/Safety**
5. **Response Length**
6. **Localization**
7. **Structure, Writing Style, and Tone**

Ranking Scale: Use the Preference Ranking Likert Scale to determine how the responses compare to each other. If Response 1 scored higher in most dimensions, it should receive a higher preference rating.

Writing Justifications for Preference Ranking

In this step, you will provide a justification for your preference ranking by explaining why one response was better than the other.

1. You do not need to quote the response while writing the justifications (simply pointing out the evidence and claims from the response are enough).
2. Stick to the point - Mention the Final Conclusion, Claim and Provide Evidence. Do not include areas that were good in both the responses, only focus on the factors that differentiate the two responses substantially.
3. Do not include or discuss dimensions which are not key differentiating factors between the two responses.
4. Do not include or discuss dimensions that do not have any issues, if their rating has No Issues, then they don't need a mention.
5. Pay closer attention to the depth and completeness of the response, over writing style, response length, and formatting while explaining and choosing which response is better.
6. Avoid flowery language, and over-explaining what is obviously reflected in individual response ratings. Redundant, irrelevant details will make your justification poor.

Example:
"Response 1 is much better than Response 2. This is because Response 2 has a major accuracy error. Response 2 states that Quentin Tarantino's Pulp Fiction (1994) grossed over $1B, but it has only grossed $212,891,760 worldwide. This critical factual error makes Response 1 the better of the two.

Step 4 - Response Rewrites

In this step, the preferred response should be rewritten to fix the issues you identified in the earlier rating. Prioritize fixing the most important issues (e.g. Truthfulness). **PLEASE DO NOT CHANGE MORE THAN YOU NEED TO.** You should also limit your rewrite to 20 minutes MAX.

Rewrite Step by Step Guide:

1. Read the prompt again.
2. Fix any critical issues you identified in the prior response rating. Again, please do not change more than you need to. Some examples of fixes include:
   a. Correcting inaccurate statements.
   b. Ensuring the response addresses all aspects/constraints of the prompt.
   c. Fixing spelling, grammar, and formatting mistakes.
3. Re-read the prompt.
4. Double check all issues are resolved and new issues have not been introduced.

Here are some examples:

- Truthfulness:
  - Example:
    - "..Moonrise Kingdom, *which grossed $62 million worldwide, was one of Wes Anderson's most financially successful films. The film acted as a vehicle for Bruce WIllis and Jared Gilman, and made more money than* The Royal Tenenbaums*…"*
  - Facts:
    - *Moonrise Kingdom* grossed $68 million worldwide
    - *Moonrise Kingdom* is the third most financially successful Wes Anderson films
    - *Moonrise Kingdom* stars Bruce WIllis and Jared Gilman
    - *Moonrise Kingdom* made less money than *The Royal Tenenbaums*, the next most successful movie behind it was *Fantastic Mr. Fox*
  - Rewrite:

- Instruction Following:
    - Example:
        - Prompt: Can you tell me about 5 fun and awesome inventions by women from the last 50 years that aren't electronic?
        - Text: "*...4. Silly Bandz - Not all inventions have to be life-changing; some can simply bring joy! Silly Bandz are colorful, rubber band-like bracelets shaped into various fun figures. Kids can wear them and trade them with friends. They were invented by BCP Imports, founded by Robert Croak. ...*"
    - Explanation:
        - Regardless of whether or not this fact is true, the prompt asked for inventions by women. This result is incompatible because the response strongly implies that it was invented by a man.
        - **IMPORTANT:** When doing this rewrite, you cannot simply delete the result, the prompt asks for **5** results, you **must** replace it with something else
    - Rewrite:
        - "*... 4. Geobond- Some inventions are cool and exciting because of how they change the physical world we live in! Geobond was invented by the sculptor Patricia Billings in the late 1970s. Geobond can be added to concrete and gypsum to create a fire-proof non-toxic material which is still really tough! ...*"

Appendix

Confusion between Dimensions

We already talked about Truthfulness vs Instruction Following but here are some more explanations:

1. **Instruction Following vs Structure, Writing, and Tone**
    a. Structure, Writing, and Tone has to do with the organization of the response, visual presentation, and text formatting while Instruction Following has to do with the asks and constraints of the prompt.
    b. There is overlap between Structure, Writing, and Tone and Instruction Following when the prompt has constraints around aspects of the dimension, usually tone (e.g. summarize in the tone of a 5th grader) or specific formatting instructions. You can ding for BOTH Structure, Writing, and Tone if the response does not follow the requests.
2. **Response Length vs Structure, Writing, and Tone**
    a. Response Length has to do with whether the response is repetitive or focused, and if it contains relevant supporting information that adds value to the main points. This relates more to the content of the response.
    b. Structure, Writing, and Tone has to do with the organization of the response, visual presentation, and text formatting.
3. **Localization vs Truthfulness**
    a. Localization has two parts: A) Ensuring key local elements (e.g., customs, regulations, or names) are accurate B) Ensuring the translation/language/punctuation/grammar/spelling that is specific for that locale is accurate. A good test is: Does the model's response content and text feel authentic and relevant to a specific region or country? Would a person from your locale think that the chatbot

English - US task, it's both a Localization and Truthfulness issue!

i. To be very clear, missing the actual date for United States Independence Day would not be a localization error for a Japanese task. The United States Independence Day is not local context/info for Japanese people. This would only be a Truthfulness error.

Mixed Language Guidelines

No reference text

Prompts should be written in the local language assigned. However, sometimes the model makes mistakes when generating the response. Here is how to grade.

Prompt language	Response language	Error
In locale language (e.g. Japanese) *A prompt in one language implies the response should be in same language*	Few words/phrases in another language that fit the context, i.e. 9/10 locale speakers would rate as normal usage in their language. (Rest of the response is in the target locale)	✅ **No error**
	Few words/phrases in another language that are out of context, i.e. 9/10 locale speakers would rate as abnormal/weird usage of the words/phrases in their language. (Rest of the response is in the target locale)	❌ **Localization minor issue**
	1 sentence to 1 paragraph in another language. (Rest of the response is in the target locale)	❌ **Localization minor/major issue, depending on CB judgement** ❌ **IF minor issue**
	Majority of the response is in another language (> 1 paragraph to entire response).	❌ **Localization major issue** ❌ **IF major issue**

Reference text

Prompt language	Response language	Error
Prompt in locale language **BUT reference text is in another language**	Entire Response is in the target locale (expected model behaviour).	✅ **No error**
	Few words/phrases in other language that fit the context, i.e. 9/10 locale speakers would rate as normal usage in their language. (Rest of the response is in the target locale)	✅ **No error**
	Few words/phrases in other language that are out of context, i.e. 9/10 locale speakers would rate as abnormal/weird usage of the words/phrases in their language. (Rest of the response is in the target locale)	❌ **Localization minor issue**
	1 sentence to 1 paragraph in other language. (Rest of the response is in the target locale)	❌ **Localization minor/major issue, depending on CB judgement** ❌ **IF minor error**
	Majority of the response is in other language (> 1 paragraph to entire response)	❌ **Localization major issue** ❌ **IF major issue**

Note: We don't encourage adding a reference text in another language

In locale language: Ex: 한국의 전통 음식에 대해 설명해주세요 (ko_KR)	A few words in english that makes sense for the locale customs Example Response: 한국의 전통 음식은 다양하고 맛있습니다. 대표적인 음식으로는 김치, 불고기, 비빔밥 등이 있습니다. 김치는 발효된 채소로, 특히 배추와 고추가루로 만든 것이 일반적입니다. 불고기는 양념된 고기를 구워서 먹는 요리로, 쇠고기나 돼지고기를 사용합니다. 비빔밥은 밥에 다양한 채소와 고기, 계란을 넣고 고추장으로 맛을 낸 음식입니다. 한국에서는 이런 음식들이 "comfort food"로 여겨지며, 특별한 날이나 일상에서 자주 먹습니다.	✅ **No error** Response: Although this is an English term ("comfort food"), it doesn't disrupt the understanding of the task and doesn't undermine the overall task, as the primary content is still in Korean and the usage of the English word is in context.
	A few words in English that are out of place: **Example Response:** 서울의 강남역은 bustling 한 곳으로 유명합니다. 이 지역에는 많은 restaurants 와 cafes 가 있어 항상 사람들이 많습니다. 특히, 주말이면 crowds 가 매우 많아서 걸어다니기 어려울 정도입니다. 강남의 쇼핑몰에는 brand-name 제품들이 가득 차 있어, 쇼핑을 좋아하는 사람들에게는 paradise 와 같습니다. 또한, 이곳은 nightlife 가 활성화되어 있어, 밤늦게까지 "bars 와 clubs 가 열려 있습니다.	❌ **Localization minor issue** **Response:** In this response, English words like "bustling," "restaurants," "cafes," "crowds," "brand-name," "paradise," "nightlife," and "bars" are inserted into otherwise Korean sentences. While many of these terms are common in modern conversations, especially in big cities like Seoul, their use is not fully appropriate in a Korean context, especially if there are equivalent Korean words available.

Major vs. minor errors

Instruction Following

Key questions to ask:
Does the response successfully do what is asked?
Does the response follow the format, length, tone, exclusions, or other constraints explicitly mentioned in the prompt?

- Major issues mean a response is **totally wrong and not at all useful for the user's explicit intention**.
- No issues mean a response is generally good and can include 1-2 aspects where implicit or unclear asks were not addressed or only partially addressed
- Minor issues mean a response is generally following instructions but misses on a few instructions BUT this doesn't make the response impossible to use

Prompt	No issue	Minor Issue	Major issue
Overview	*Response is useful with only minor deviations*	*Response is imperfect, but overall still useful*	• *Response is totally unhelpful* • _Response is a punt not driven by safety*_
Give me 5 cafes in my neighborhood.	• 5 cafes ◦ All in business and in the neighborhood • 5 real cafes in the right area, all are real and in business, ◦ Opening time is wrong for 2 cafes	• 5 real cafes in the right area ◦ All are / were real ◦ 2 are closed permanently Why this is a minor issue: *The information provided here clearly has issues, but it still follows key requirements in*	• 5 cafes ◦ 3 cafes are hallucinated (cafe names are made up) ◦ 2 are closed • Provides cafes in a whole other city • Provides establishments that cannot be

			Why this is a major issue: *The information provided here clearly does not follow important requirements highlighted in the prompt.*
Summarize this document in 5 bullet points	• Summary in 5 bullet points • Acceptable to paraphrase from text • Acceptable to describe concepts / synthesize meaning that is implied but not explicit in text	• Summarizes key points, hallucinates minor details not key to the meaning • Summarizes in 5 sentences but not in bullet points • Summarizes in 4 or 6 bullet points (+/-1) *Bullet points are formatting / length constraints rather than the core requirement, so follow +/- 1 guidance*	• Substantially hallucinates • Gets key details core to the meaning so wrong the summary is not at all useful • Summary in 3 or 7 bullet points
“What are the pros and cons of remote work?"	• Provides pros and cons in some format • Acceptable if some pros and cons are not clear-cut / are subjective	• Response lists pros and cons as single-word bulleted entries with no description	• Writes an essay about something totally unrelated • Describes remote work but does not in any way includes pros and cons

Truthfulness

Key questions to ask:
Are factual claims in the response grounded in reference text (if present), or verified as accurate?
Does the model accurately deliver the stated core requested?

- Major issues mean a response is **not all accurate to the core request of the prompt making it not at all useful for the user’s explicit intention**.
- No issues mean a response is generally factually correct and accurate in context of the core requirements
- Minor issues mean a response is generally accurate with a minor inaccuracies or inaccuracies on facts peripheral/*ancillary to* the core request BUT this doesn’t make the response impossible to use

Note: The below are truthfulness issues ONLY if the model does NOT restate the core requirement and provide accurate information based on that restatement

Prompt	No issue	Minor Issue	Major issue
Overview	*Response is correct in factual claims and contextual accuracy (provides what it states or what the prompt requests)*	*Response has a few inaccuracies but is truthful enough to still be useful to the user*	*Response has factual or contextual accuracies major enough to make the response not at all useful to the user*
Give me 5 cafes open in my neighborhood.	• 5 cafes ○ Cafes are ‘open’ as in in business ○ If provided, facts like hours of operations and addresses are accurate	• 5 cafes ○ Hours of operation were listed incorrectly • 5 cafes ○ 2 of the addresses were listed incorrectly Why this is a minor issue: *Some of the information provided is not factual, but the details core to the*	• 5 cafes ○ 2 are hallucinated ○ 1 is closed • Provides something totally wrong • Provides cafes in a whole other city

	• Acceptable to describe concepts / synthesize meaning that is implied but not explicit in text	details not key to the meaning	meaning so wrong the summary is not at all useful
“What are the pros and cons of remote work?"	• Provides pros and cons in some format • Acceptable if some pros and cons are not clear-cut / are subjective	• Response lists pros and cons but provides an incorrect statistic in support of one of the pros	• Response is about something totally unrelated • Describes remote work but does not in any way includes pros and cons