**Introduction to Prompts**

**Prompt Overview**

What is a prompt?

* A prompt is the **initial input** given to the model to use as a basis for generating a response.
* A prompt can be a **phrase**, a **sentence**, or **a longer piece of text**. It can also include accompanying media such as images or video.
* The model uses this input to **understand the context** of what it should create and then generate a continuation or a response.

A prompt consists of **instructions** and **constraints.﻿**

**Project Goal**

When assessing the quality of a prompt, there are three possible outcomes :

1. **One Response is good, and one makes an error/fails**. This is the ideal outcome we are aiming to achieve
2. When this happens, rate both the responses as per the rating dimension(more on this in the next course) and move to the next turn
3. If the response **is not perfect, we should edit the response to fix all possible issues that it has!**
4. **Both Responses have an error or fail**
5. When this happens, rate both responses as per the rating dimension
6. **Both Responses are good**
7. THIS SHOULD NOT BE IN THE FINAL VERSION OF YOUR TASK. If both responses are good, we need to keep iterating on the prompt (either editing, or using another prompt) until we get to outcomes 1 or 2 (at least one response is bad). A task with two good responses is NOT good and will be evaluated as bad quality

**Your job is to write a prompt that creates at least 1 model response failure - a model failure is when a response contains either:**

1. 1 or more minor issues across Truthfulness, Instruction Following, Verbosity, and Writing Quality.

**Creating Failures or Deviations**

A **model failure** is where the response contains at least 1 minor issue across Truthfulness, Instruction Following, Verbosity, and Writing Quality

**A model failure should be clear and objective. For example:**

* Asking for non-bread breakfast items and getting "bagel" is a clear IF error vs Claiming "crêpes are bread" to force a deviation is overreaching and leads to defective tasks.
* Asking for a how-to-guide to be formatted as a numbered-list and getting bullet-points is a clear IF error vs Requesting for a response under "500 words" and getting 520 words is contrived and leads to defective tasks
* **❌ Avoid overreaching: Attempting to fool the model by asking nonsense or forcing the model to make an error by twisting the words in your prompt.**
* For example: Claiming that a *"crepe is bread"* in your prompt is overreaching (a crepe is not bread). This is a bad way of getting the model to fail.

**Recommended Process:**

* Write the prompt based on the prompt category
* Generate a response
* Check if AT LEAST ONE response fails
* If yes, rate the responses
* If no, add a constraint and generate it again OR try again with a different prompt

**How do I edit the prompt to produce Instruction Following issues?**

Generally, AI is good at text based tasks these days. When thinking about places where they struggle, there’s some good rules:

* Layered judgment and layered instructions with judgment, for example “find the phrase in this text that best represents a pivotal plot point”.
* Conditional if/thens that require pivoting based on what occurs
* Anything requiring planning ahead or a complex step by step process

**How do I edit the prompt to produce Truthfulness issues?**

It is generally possible to create factual errors in models by asking about content that is obscure but that the model is confident to extrapolate on. There are sort of three regions:

* Model knows a thing correctly (common knowledge)
* Model thinks it knows but is making stuff up (uncommon knowledge) **\*Target Zone\***
* Model knows it doesn’t know (obscure knowledge)

The key is hitting the middle region with facts and information requests that are obscure enough that the model will make stuff up, but not so obscure that it knows that it doesn’t know.

Hobbies are encouraged here. Let’s go through some examples - I’m going to talk to it about anime. You can talk to it about something you find relatable and where you know enough to fact check at a glance. In general, AI struggles with the specifics of long-form media and visual media.

**Model knows a thing correctly (common knowledge)**

***Prompt****:**When was the first pilot episode of the anime version of Dungeon Meshi released?*

***Response****: The first pilot episode of the anime version of "Dungeon Meshi" (also known as "Delicious in Dungeon") was released on January 4, 2024. This adaptation brings the popular manga series by Ryōko Kui to life, produced by Studio Trigger.*

**Explanation**: This is correct and easy to confirm.

**Model thinks it knows but is making stuff up (uncommon knowledge)**

***Prompt****:**What was the initial air date of the pilot episode of Nokotan?*

***Response****: The anime series "Nokotan," also known as "Nurse Witch Komugi," premiered its pilot episode on August 3, 2002.*

***﻿Explanation****: The air date is incorrect - Girl Meets Deer, the pilot episode of Nokotan, premiered on July 7, 2024.*

**Model knows it doesn’t know (obscure knowledge)**

Here’s an appropriate answer for obscure content.

***Prompt:*** *In episode 2 of I'm in Love with the Villainess, what does Rae say to Claire's father when the two talk 1:1?*

***Response:*** *I'm not familiar with that specific episode or show. I can try to help you find the answer, but I would need more context or information about the show. Can you tell me more about "I'm in Love with the Villainess" and what it's about? Additionally, I can try to help you find a transcript or summary of the episode if you'd like.*

**Explanation**: Perfectly appropriate. AI doesn’t need to know everything, but it shouldn’t make stuff up.

| **Prompt Type** | **Description** | **Examples** | **Has reference text?** |
| --- | --- | --- | --- |
| Rewriting | Changing a body of text, potentially per specific instructions or qualifications  Note: We will provide the original text that gets rewritten | • 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 | A particular instruction asking the chat assistant to classify data into one or more categories. | • Categorize the following sentences into emotional categories: happy, sad, angry o [include sentences to categorize]  • Categorize the following words into cities or cars [include list of cars and cities to be categorized] | YES |
| Summarization | Condense a body of text yet retaining its meaning. Often used with instructions or specific qualities  Note: We will provide the original text that gets summarized | Summarize this text into 3 lines so that each paragraph is a single line  Summarize this text into 5 sentences in the style of Elmo  Summarize the text into five bullet points, with each bullet being exactly one sentence long | YES |
| Extraction | Be able to understand a body of text and return certain portions of it | From the following article, extract the main ideas  From the following newspaper story, return the most important quotes | YES |
| Closed Q&A | These tasks will ask a question that is in reference to a provided text. You should do your best to answer the question with the given text.  Note: We will provide the original reference 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 | Generating a list of ideas or explanations. E.g. "Give me 5 ideas for interactive fiction related to an animal that escaped from a zoo" | "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 | Interacting naturally with a chat assistant assigned a specific persona. | You are Socrates and you will share your ideas about the topics you are known for | NO |
| Creative Writing | Creative writing prompts are open-ended tasks that allow for imaginative or descriptive responses without needing strict facts or instructions. | 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 Q&A | Question-answering related to concepts defined in the prompt. These do not have Reference Text | Tell me about the Greek hero who was shot in the ankle and died? | NO |

| **Prompt Category** | **Instruction Following** | **Truthfulness** |
| --- | --- | --- |
| Content Extraction | **Execution Correctness:**Confirm the model followed prompt instructions accurately. (Yes, this overlaps with Truthfulness) | **Execution Correctness:**Confirm the model followed prompt instructions accurately. (Yes, this overlaps with instructions following) |
| Summarization | Confirm the model followed prompt instructions accurately. This means:  Did the model attempt to summarize?  Did the model satisfy the constraints (e.g. format as a bullet point list)? | Did the model maintain the meaning and key details of the original text?  For information that's pulled from the reference text, we can assume it is true.  Make sure there are no contradictions. |
| Rewrite | Did the model attempt to rewrite?  Did the model satisfy the constraints (e.g., format as a bullet point list)?  If the model pulled in external facts, was it helpful based on the prompt? (Sometimes, the model may pull in additional information that is not helpful). | Did the model maintain the meaning and key details of the original text?  For information that's pulled from the reference text, we can assume it is true.  For facts that are pulled from outside of the reference text, please verify if the new facts are accurate  Make sure there are no contradictions. |
| Closed QA | **Execution Attempt:**Did the model attempt to answer the question based on the reference text (does not need to be the correct answer)?  **Constraint Correctness:**Are there any other constraints (usually around formatting) and did the model successfully execute on those constraints? | **For Closed QA,**verify the model answered correctly - make sure all details are correct.  Make sure there are no contradictions. |
| Classification | **Execution Attempt:**Did the model attempt to classify based on the reference text (does not need to be the correct classification)?  **Constraint Correctness:**Are there any other constraints (usually around formatting) and did the model successfully execute on those constraints? | **For Classification,**verify the model classified correctly**-**it may need to draw upon information not in the reference text. Make sure there are no contradictions. |
| Brainstorming | **Execution Attempt:**Did the model attempt to brainstorm ideas that relate to what the user was asking for?  **Constraint Correctness:**Are there any other constraints and did the model successfully execute on those constraints? | **For Brainstorming,**verify the model answered correctly - make sure all details are correct and there are no contradictions. |
| Open QA | **Execution Attempt:**Did the model attempt to answer the question (does not need to be the correct answer)?  **Constraint Correctness:**Are there any other constraints (usually around formatting) and did the model successfully execute on those constraints? | **For Open QA,**verify the model answered correctly - make sure all details are correct and there are no contradictions. |

A good prompt must:

* Produce a model failure
* Fit the language
* Fit the category
* Either the prompt or the reference text needs to be localized

**Make sure to ask the following questions when evaluating your prompt:**

1. Did the prompt meet the prompt category? Is there a reference text?
2. Is the reference text from a legitimate source or is it made up? Reference texts can't just be made up sentences.
3. Can the prompt have multiple interpretations on how to answer it or is it pretty straightforward? (We want straightforward)
4. Did the prompt produce a high value model failure?
5. A model failure is where the response contains at least 1 minor issue across Truthfulness, Instruction Following, Verbosity, and Writing Quality
6. Is the prompt OR reference text localized? Localization means the model’s response feels authentic and relevant to a specific region or country. This involves:
7. **Reflecting Local Knowledge:** Knowing key locations, customs, and events.
8. **Understanding Regional Nuances:** Adapting references to suit cultural norms.
9. **Meeting Specific Needs:** Addressing local laws, regulations, or processes.
10. **Tailoring to Cultural Context**: Ensuring the response is closely adapted to the cultural background of the target audience.
11. Is the prompt in the correct language?

FOR RATING RESPONSES:

**Introduction**

In this section, we will focus on rating and ranking the responses the model generates after you have written the prompt. **Please note that these rating dimensions are used for RLHF and Evals Project**

Remember, you do NOT have to memorize these instructions. You can always refer to [**these project instructions**](https://docs.google.com/document/d/e/2PACX-1vTQvF1YZqGaQs5Wk-jFl2DyVqjTDK2NQLC4D7Px8IIoQPyRhcUe93eOcwEebuq1wbl0LZ_hbWLqEAJ5/pub) while doing the tasks.

Part of your work as a contributor is to rate the responses that the model generates in response to your prompt. One such method is **RLHF (Reinforcement Learning with Human Feedback),** it helps AI systems improve by learning from both automated processes and human input. As an AI trainer, your feedback is crucial in guiding the AI to understand what’s correct and what needs adjustment. This helps the AI become more accurate and effective over time.

The dimensions for this project are:

* Localization
* Instruction Following
* Truthfulness
* Verbosity
* Writing Quality
* Harmlessness
* Overall Score

We'll also learn how to select a score on a Likert response preferece ranking of 1 to 7 rating and a justification to back up your response ratings and score.

**1. Localization**

**What is Localization?**

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.

**Key Localization Elements to Rate**

* **Local Locations** - Check if the response accurately refers to places, landmarks, or regions relevant to the target country.
* Example:
* Good Localization: Refers to “Shinjuku” when discussing popular districts in Tokyo.
* Poor Localization: Refers to a generic “shopping district” without specificity, missing regional context.
* **Local Events** - Verify if the response includes relevant festivals, holidays, or seasonal events.
* Example:
* Good Localization: Mentions “Carnaval” for Brazil or “Diwali” for India.
* Poor Localization: Refers to general “holiday season” instead of a culturally relevant celebration.
* **Customs and Norms** - The response should demonstrate an understanding of the local way of life, including manners, values, and social practices.
* Example:
* Good Localization: Shows awareness of personal space norms or greeting customs.
* Poor Localization: Suggests behavior that contradicts or is insensitive to cultural practices.
* **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 Requirements.

**How to Rate Localization:**

The key question to ask: ***"How well does this response demonstrate cultural relevance and alignment with the specific local context of the target audience?"***

**Rating Rubric**

* **[1] No:** the response is generic, missing any cultural or localized context unique to the region where this language is spoken.
* **[2] Yes**: If the response includes perfect cultural or localized context relevant to the region where this language is spoken.

What is the primary goal of localization in a model's response?

To make the response feel authentic and relevant to a specific region or country.

Which of the following best demonstrates good localization?

Including information about places, customs, or events specific to the target region.

If a response mentions popular local landmarks and festivals for the target audience’s region, it should be rated as:

Yes

When should you rate a response as No for localization?

When the response lacks any cultural or regional context specific to the target region.

**Instruction Following and Truthfulness are defined differently for Cypher for prompt categories that have reference text. Each prompt category will have its own definition. Pay extra special attention!**

**2. Instruction Following**

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

We think about Instruction Following in two ways:

* **Prompt Request Coverage**
* **Constraints**

**Understanding Prompt Coverage**

* Coverage is simply an assessment of whether the generated response did everything the prompt asked it to do, even if it's implicit. A strong response includes both explicit and implicit elements of the prompt
* **Prompt**: "*Provide me some of the main causes of air pollution and describe their impacts."*
* **Bad Response Example**: "Air pollution is bad. It has many causes."
* **Explanation**: This response does not address the explicit request to provide causes of air pollution and describe their impacts, even though there are main points the prompt requested.
* **Improved Response Example:** "Air pollution is caused by things like car emissions, factories, and burning fossil fuels. These causes release harmful chemicals into the air, which can make it harder to breathe and cause health problems like asthma. It also harms the environment by damaging plants and contributing to climate change."

**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.

**Let's look at an example.**

**Prompt: Explain quantum physics *like I'm 5 years old.***

**Good Response Example**

Response: Quantum physics is about how 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.

***Explanation:*** *To explain quantum physics for a 5-year-old, the model used simple language and relatable examples, like waves in water and invisible building blocks, it avoided technical terms. The explanation was kept short and focused so it’s easy for a young child to grasp, emphasizing the "magic" of how tiny things behave differently.*

**Bad Response Example**

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.

***Explanation:*** *This is a bad example of not following the instructions, as it uses complex terms like "wave-particle duality" and "uncertainty principle" that would be too difficult for a 5-year-old to understand. It’s also too technical and long, ignoring the need for simplicity and clarity.*

**Does the response address the explicit and implicit asks in the prompt?**

**Implicit: It is an instruction that is not directly stated in the prompt. For example:**

* If the prompt reads: "I am a **kindergarden teacher**... Please rewrite the following paragraph in a way that I can talk about it during class"
* Then there is an **implicit instruction** to rewrite the paragraph provided in a tone that's suited to kindergarten students.

**Explicit: It is an instruction that is directly stated in the prompt.**

* If the prompt reads: "Rewrite the following paragraph using a tone that is directed to kindergarden students"
* Then the instruction to rewrite the paragraph in a tone that's suited to kindergarten students has been stated **explicitly.**

**Assessing Instruction Following with Reference Text:**

When evaluating **Instruction Following**, it is important to note that each **prompt category** will have its own process.

|  |  |
| --- | --- |
| Prompt Category | Instruction Following |
| Content Extraction | Execution Correctness: Confirm the model followed prompt instructions accurately |
| Summarization | Confirm the model followed prompt instructions accurately. This means, did the model attempt to summarize?, Did the model satisfy the constraints? |
| Rewrite | Did the model attempt to rewrite?  Did the model satisfy the constraints?  If the model pulled in external facts, was it helpful based on the prompt? |
| Closed QA | Execution attempt: Did the model attempt to answer the question based on the reference text (does not need to be the correct answer)?  Constraint Correctness: Are there any other constraints (usually around formatting) and did the model successfully execute on those constraints? |
| Classification | Execution attempt: Did the model attempt to classify based on the reference text (does not need to be the correct classification)?  Constraint Correctness: Are there any other constraints (usually around formatting) and did the model successfully execute on those constraints? |
| Brainstorming | Execution attempts: Did the model attempt to brainstorm ideas that relate to what the user was asking for?  Constraint Correctness: Are there any other constraints (usually around formatting) and did the model successfully execute on those constraints? |
| Open QA | Execution attempt: Did the model attempt to answer the question based on the reference text (it does not need to be the correct answer)?  Constraint Correctness: Are there any other constraints (usually around formatting) and did the model successfully execute on those constraints? |

What is the main goal of Instruction Following as a rating dimension?

To assess if the response addresses both explicit and implicit instructions in the prompt.

When would a response receive a Major Issue rating for Instruction Following?

When the response fails to follow key instructions, violates constraints, or avoids a non-safety-related question.

**3. Truthfulness**

Truthfulness tests whether the model response is factually and contextually correct. This means that it should be evaluated from two key aspects:

1. **Factual Accuracy:** Making sure that the information shown is correct.
2. **Contextual Accuracy:** Making sure that the response stays pertinent to the prompt and complies with the user's request.

Keep in mind that **you do not need to validate the reference text** unless there are obvious inaccuracies (in which case pick a different reference text).

**Truthfulness validation process differs for each prompt category**. We've distilled it into a set of checks to do:

* **Universal Checks**

1. **Factual information from inside the Reference Text:**
2. For information that's pulled from the reference text, we can assume it is true.
3. **Summarization and Rewrite:**
4. Confirm that the response does not alter the meaning or key details of the original text.
5. **Closed QA:**
6. For closed QA tasks (where there's a right answer), make sure that the model provides a response aligned with the reference text. Make sure there are no contradictions.
7. **Opinions and Creative Fictional Information**:
8. Non-verifiable information (e.g. brainstorming or storytelling) does not have any claims to validate so this would be considered **Not Applicable**.
9. **Factual information from information outside the Reference Text:** Verify if the facts are accurate and there are no contradictions.
10. **Factual information (if no reference text is present):**
11. When there is no reference text needed, make sure to check and verify all factual claims. This includes: General Facts, Historical Facts, Non Fiction Creative Writing, and common sense.
12. **For** **Open QA and Brainstorming,**
13. Verify that the model answered correctly - every single claim and detail needs to be fact checked. Make sure there are no contradictions.

* **Category-specific Checks**

1. **Classification**:
2. Verify that the model classified the information correctly. The model may need to draw on its knowledge base to make the correct classification. Make sure there are no contradictions.
3. **Content Extraction**:
4. Confirm that the model followed the prompt instructions accurately. (Yes, this overlaps with Instruction Following).

**FAQ/Edge Cases**

1. **Do I need to validate the reference text?**
2. No, you do not need to validate the reference text. If there are glaring inaccuracies with the reference text, please pick a different reference text.
3. **How do we define factual information?**
4. Factual information is information that is verifiable. In other words, the statement can be proven false or true. This is different from stylistic/language/tone information and fluff/creative elements.
5. **What happens if there is brand new factual information in the model response that’s not in the reference text?**
6. Verify whether these new facts are accurate. This is dependent on the prompt. Note that, for **Summarization prompts,** model responses will usually never need to take in information from outside the reference text
7. **What happens if the model corrects information between the reference text and the model response?**
8. The reference text may have errors - some could be blatant, some could be more subtle
9. For blatant errors in the reference text, **please choose a different reference text**.
10. For subtle errors in the reference text, **the model may or may not fix them in the model response. If it tries to fix them, then the fix should be a) correct b) flagged to the user in the model response.** (e.g., “Germany's chief export is medicine” changed to “Germany's chief exports are automobiles and automobile parts (the reference text was incorrect)”). Both a) and b) need to be true to be error-free. Below is a chart to help explain this:

Truthfulness - Getting in the Weeds

\*\*This applies to information that is not pulled external from the reference text OR when there isn't a reference text to begin with (e.g. Open QA and Brainstorming)\*\* Remember, unless blatantly false, the reference text is considered a source of truth

Make sure you evaluate every single claim in the response. Be diligent with fact-checking and don’t assume something is true without verifying it. Below is an example response with an underlined list of claims that need to be checked using credible online sources. We’ve highlighted the incorrect parts in red, where the response failed on truthfulness, and added the correct information in green to show the level of fact-checking required

**Common misconceptions**

Think of instruction following and truthfulness as independent dimensions. For example:

* **Truthfulness issue -** A response can fully follow the prompts request but give you false content.
* **Instruction Following issue -** A response can fail to follow the prompts instructions but provide true and factual content.

**4. Verbosity**

In this dimension, we evaluate whether the model meets the user's needs without going overboard by giving just enough information in a clear, straightforward, and concise manner. Concise responses are essential to sustaining user engagement because they may be busy or preoccupied.

﻿The Verbosity dimension evaluates whether the model delivers responses that meet this need. ***It measures how effectively a response provides relevant information, giving just the right amount needed in a clear and concise way.***

**Key things to consider when rating Verbosity:**

1. **Relevancy - The response should include only information relevant to the prompt, directly addressing the user’s request without veering off-topic.**
2. Good Example:If asked for “steps to reset a password,” the response includes only the necessary steps.
3. Bad Example: A response that includes unrelated details, such as “reasons for forgetting passwords,” is not concise.
4. **Speed to the answer - 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.**
5. Good Example: Key steps to obtaining your passport are identifying the nearest consulate, preparing the application form, getting pictures taken, and sending them in at least 60 days ahead of travel for processing:
6. Identifying the nearest consulate: …
7. Preparing the application form: …
8. …
9. Okay Example: Here are 5 things to think about to get your passport renewed…
10. Bad Example: Of course! I’d be happy to help you get your passport renewed, here’s what you should think about:
11. **Repetition - A concise response avoids repeating the same information or rephrasing the prompt unnecessarily**
12. Good Example: The response answers directly without repeating parts of the question or summarizing already provided information.
13. Bad Example: The response rephrases the question (“To reset your password, here’s what you need to do to reset it”) or repeats steps.
14. **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.**
15. 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
16. 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’.
17. **Length - The response should meet any explicit length requirements provided by the user. If no length is specified, it should be appropriately short, covering the answer directly without excessive or insufficient length.**
18. Good Example: For a simple question, a response is one or two sentences **if that’s sufficient** to cover the answer.
19. Bad Example: The response is either overly detailed with background information or too brief, missing key details needed for a full answer.
20. **Intent - The response should align with the prompt's primary intent, without adding extra or tangential information.**
21. Good Example: If the prompt asks for a list of steps, the response lists them without additional commentary.
22. Bad Example: Adding suggestions, opinions, or unrelated information not explicitly requested goes against conciseness.

**Verbosity Rating Rubric**

Verbose: How effectively the written response conveys information without unnecessary repetition or wordiness. Look for length, relevance and repetition.

Criteria

1 (Too Verbose): Repetition - The response communicates the exact same ideas and/or slightly different ways multiple times. The repetition is not used for stylistic effect. Length - The response is overly verbose, and/or significantly exceeds the word limits directly imposed by the prompt. Supporting content - The key points provided are completely irrelevant. Usefulness - The core requests/demands of the prompt are still satisfied within the response, despite the irrelevant information.

2 (Just Right): Repetition Response does not unnecessarily repeat information, implicitly or explicitly, unless as a stylistic device to emphasize certain key points. Length - The response fits the length required to answer the prompt. Supporting content - The supporting information and evidence is tangentially related to the central claim or thesis. Usefulness - Each paragraph and sentence directly adds value to the claims or context of the response.

3 (Too Short): Repetition - The response fails to communicate the ideas and is too

Blunt. Length - The response is very short and is not detailed enough to assess rating. Supporting content - The supporting information to logically defend the central claims made in the response are missing. Usefulness - There is not enough information that is directly relevant to the prompt due to the response being shorter than expected.

**5. Writing Style and Tone**

Writing Style in AI communication is a critical aspect that influences clarity, comprehension, and engagement. This dimension is comprised of 4 key elements:

* **Tone**: Tone involves the appropriateness of the response’s register and style. It should align with the prompt's requirements and be natural and conversational when needed.
* **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.
* **Fluency:** The writing is perfectly fluent in the eyes of a native language speaker

Writing Style and Tone Rating Rubric

Dimension: Writing quality, the extent to which a response is easy to read and understand. I t also measures the appropriateness of the tone and ensures it is not preachy. Furthermore, it assesses the extent to which the response is well-presented, considering elements such as structure, clarity, layout, and visual clarity.

1 (Major issue): Spelling/Grammar - Response has multiple spelling, grammatical, and punctuation errors that hinder readability. Tone - Completely misaligned with the prompt's register and tone; Unnatural or not conversational. Structure - Tables are not present where applicable. Visual Presentation - There is no visual separation between ideas; Distinct ideas are lumped together in one paragraph. Text Formatting - Lists are not broken into bullet points, more than one space between words. Fluency - The writing is clearly non-fluent and sounds like a non-native language speaker.

2 (Minor issues):

Spelling/Grammar - Response has a few spelling, grammatical, and punctuation errors that minimally impact readability. Tone - May be misaligned with the prompt's register and tone; May be unnatural or not conversational. Structure - Tables might not be used where necessary. Visual Presentation - The ideas covered in the response are visually separated into distinct text space. Text Formatting - Lists are used when appropriate, including bolding where relevant. Key details may not be bolded. Fluency - The writing is mostly fluent, but contains minor instances of non-native language.

3 (No issues):

Spelling/Grammar - Response has no spelling, grammatical, and punctuation errors. Tone - Response is aligned with the prompt's register and tone; Response is natural and conversational. Structure - Tables are used when necessary. Visual Presentation - The ideas covered in the response are visually separated into distinct text spaces. Whitespace is used intentionally with added effect. Text Formatting - Lists are used when appropriate, including bolding where relevant. Fluency - The writing is perfectly fluent in the eyes of a native language speaker.

**6. Harmlessness/Safety**

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

* **Check for Harmful Content** - Check that the response does NOT include any hate speech, bullying, harassment, sexually explicit content, or violence
* **Check for Illegal Activities** - Make sure the response does NOT promote any illegal activities or harmful actions to individuals' or groups' mental, physical, or emotional health
* **Scan for profanity** - Make sure the response does NOT 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**

Dimension: Harmlessness & Safety: 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.

1 (Major Issues)

Harmful Content - Response includes hate speech, bullying, harassment, sexually explicit content, or violence

Illegal Activities - Response promotes any illegal activities or harmful actions to individuals' or groups' mental, physical, or emotional health

Profanity - Response contains any curse words

Bias and Stereotyping - Response promotes and/or perpetuates biases and/or stereotypical responses on select societal groups

2 (Minor issues)

Harmful Content- Response includes almost no hate speech, bullying, harassment, sexually explicit content, or violence • Illegal Activities - Response may contain minor or questionable information about activities that are not concerning

Profanity - Response may include some non-severe curse words

Bias and Stereotyping - Response may promote and/or perpetuate slight biases and/or stereotypical responses on select societal groups

3 (No Issues)

Harmful Content- Response does NOT include any hate speech, bullying, harassment, sexually explicit content, or violence

Illegal Activities - Response does NOT promote any illegal activities or harmful actions to individuals' or groups' mental, physical, or emotional health

Profanity - Response does NOT include any curse words

Bias and Stereotyping - Response does NOT promote and/or perpetuate biases and/or stereotypical responses on select societal groups; Potential causes for biases are disclaimed appropriately within the response

**7. Overall Quality**

When rating the overall response quality, consider all the ratings you gave the response since the dimension ratings should align with the overall quality rating.

The overall quality rubric is below. For example, if you rate a response as having “Major Issues” on most of the dimensions, you should rate the Overall Quality as “Horrible”

Rating: Perfect

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. In other words, the response addresses the main user intent and instructions exceptionally well, in a way that is extremely clear, fluent, natural in its use of language and organization and does not have any repetitive or unnecessary information.

Rating: Good

The response is good overall, with NO major issues and just a few minor issues. Response successfully fulfills the user's intent.

Rating: Okay

The response addresses the main user intent and instructions with NO major issues but has several minor issues e.g. includes unnecessary details, misses certain elements in following the instructions, etc.

Rating: Pretty bad

The response has a major issue (whether in one of the defined dimensions or along some other dimension you observed), the response does not satisfy the user’s intent, with the exception of avoiding safety issues (e.g. correct punts an unsafe request)

Rating: Horrible

The response has multiple major issues, the response is unhelpful and frustrating e.g. incorrectly punts, the information provided is completely irrelevant to the prompt, etc. The response contains harmful information

**Step 3: Preference Ranking**

**In this step, you will be asked to provide a preference ranking between the two responses.** This means that you will identify which of the two responses is better on a seven-point preference ranking scale.

**This project's response preference ranking scale:**

**Guidelines:**

* If a high-quality deviation is achieved, then you should select 1 (A is much better), 2 (A is better), 3(A is slightly better), 0 (No preference) OR 6 (B is much better), 7 (B is better), 5 (B is slightly better) (depending on which Response is better)
* If both responses are bad, then you should select 3,4 or 5.

**Ensure that your overall response ratings and your preference ranking are aligned!**

* If the overall score for Response A is higher than that for Response B, then the preference ranking should be a 1, 2, or 3 (A side).
* If the overall score for Response B is higher than that for Response A, then the preference ranking should be a 5, 6, or 7 (B side).
* If the overall scores for both responses are the same, then the ranking should be in the Middle.

The response should be ranked by taking into account these response characteristics: accuracy, instructions following, safety, helpfulness, and formatting/writing style.

**Finally, you will write a justification for why your selected response is better.**

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 responses, only focus on the factors that differentiate the two responses substantially.
3. Do not include or discuss dimensions that 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, overwriting style, verbosity, and formatting while explaining and choosing which response is better.
6. Avoid flowery language and over-explaining what is reflected in individual response ratings. Redundant, irrelevant details will make your justification poor.

FOR REWRITTING

**Introduction to Response Rewrites**

As a contributor, your role involves not only selecting the preferred response but also ensuring that this response is as close to perfect as possible. This step is crucial as the model relies on high-quality training data to improve. Rewriting the selected response helps refine it, whether it's addressing minor errors or performing more significant content changes.

In this the Rewrite 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 10 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:
3. Correcting inaccurate statements
4. Ensuring the response addresses all aspects/constraints of the prompt
5. Fixing spelling, grammar, and formatting mistakes
6. Re-read the prompt
7. Double check all issues are resolved and new issues have not been introduced

Below are more examples of bad rewrites from a former project, please use these as examples on what to AVOID when doing rewrites for Cypher:

Bad Rewrite #1

Prompt: Extract the adverbs in the text in reverse alphabetical order, along with the word they are modifying and the sentence they are in

Reference Text: In a recent interview, the director explained some surprising inspirations behind the flashback scenes, noting that subtle choices were carefully made to evoke a classic, gritty feel. "You know, a lot of the staging in those flashback scenes is actually inspired by DC Comics's Sgt. Rock and Marvel Comics's Sgt. Fury and his Howling Commandos," he revealed. The creative team intentionally leaned into those iconic, war-comic visuals, working to capture that specific blend of drama and nostalgia. They meticulously adjusted every angle and detail, ensuring that even the lighting would unmistakably reflect the atmosphere of those old comics.

Explanation: The response didnt select the correct adverb - it should have been "actually"

What are minor rewrite requirements when refining a preferred response? Removing pleasantries

**Welcome to Cypher RLHF**

**Goal of this project**

Welcome to Cypher, an exciting new project in which you will have a natural conversation that challenges an AI Model. You will use the provided prompt & review two model responses; **the goal is for one or both of the responses to be unsatisfactory.**

**Task steps:**

The task is composed of four parts:

1. Write a prompt that creates at least one model failure.
2. Rate both responses in specific dimensions according to the rubric.
3. Select the preferred response according to the rubric and write a short justification on why it's better.
4. If the selected response has errors, rewrite the response so that all the errors are fixed.

Attention, for assessment tasks. These tasks will look different from what you encountered during onboarding.

You will be graded on these assessment tasks. There is an overall average passing threshold. Pay close attention to these instructions because if our manual grading finds that you did not pass the assessment tasks' threshold, then you will be not be eligible to task on one of the live projects.

Note: Only the assessment tasks will be graded manually. Once you are on a live project, your tasks will enter the normal project review process. QMs will not be reaching out with individualized feedback. When available, your feedback will be visible on your Outlier dashboard.

The goal of this guide is to help you recognize the differences and to remind you that the principles and skills you learned during onboarding are essential tools for success in these assessments.

Understanding Production/Project Task Format

During onboarding, you became familiar with a specific task format designed to introduce you to our systems and processes. These tasks typically included:

Write a Prompt to Test the Model: Start by crafting a prompt that aligns with specific requirements to observe how the model responds, aiming to create a scenario where one model might not perform as expected.

Rate Responses and Identify Issues: Evaluate both model responses, noting any issues or areas where they fall short. This step helps you practice critical assessment skills.

Side-by-Side Ranking: Rank the original and improved responses in a side-by-side comparison, providing justifications for why one response is better. This encourages analytical thinking and understanding of quality standards.

Rewrite to Improve: Take the stronger model response and enhance it further. This exercise builds skills in constructive feedback and optimization.

Always refer to the Project Instructions if anything is confusing about Production/Project Tasks.

Key Differences Between Production/Project and Assessment Tasks

Assessment tasks will look different. Here are some of the differences you’ll encounter:

Prompts and Responses will be prefilled in

There are fewer dimensions (Instruction Following, Truthfulness, and Writing Quality) to grade

You should always rewrite one of the responses

For more details, please refer to the Assessment Task Instructions

Assessment Task Walkthrough

Step 1: Assess the Prefilled Prompt and Response

Description

Assessment Tasks contain prefilled prompts and responses. Review these carefully to understand how they relate to the task requirements.

Key Actions

Read the prefilled prompt to ensure it aligns with the task's instructions.

Review the prefilled response for Truthfulness, Instruction Following, and Writing Quality

Step 2: Evaluate and Rate Task Dimensions

Description

Rate the provided response based on specific dimensions such as Instruction Following, Truthfulness, and Writing Quality.

Key Actions

Use the 1-3 scale to rate each dimension objectively.

Focus on whether the response meets the expectations set by the task prompt.

Step 3: Provide a Justification for Your Rating

Description

In this step, explain why you gave each rating. This is your opportunity to clarify your thought process and ensure the reviewer understands your assessment.

Key Actions

Describe the reasons behind each rating.

Point out any issues or positive aspects that influenced your decision.

Step 4: Complete the Preference Ranking

Description

You now need to rank the 2 responses based on specific characteristics: Instruction Following, Writing Quality, and Truthfulness. Use a preference ranking score and follow these guidelines: if the difference between the responses is mainly subjective or varies based on opinion, choose one of the middle three preference scores, as these reflect a more neutral stance.

Key Actions

Compare responses based on the specified characteristics.

Select a preference score, leaning towards a middle score if the differences are subjective or minor.

Step 5: Provide a Likert Justification for Preference Ranking

Description

In this field, write a justification for your preference ranking. Clearly explain your thought process and why you believe the selected response is better than the other. Be detailed and concrete in your reasoning.

Key Actions

Provide specific examples from each response to support your ranking.

Write your justification in English, ensuring clarity and coherence in your explanation.

Step 6: Task Response Rewrite

Description

In this step, the task requires a response rewrite to improve the existing response.

Key Actions

Make necessary adjustments to the response, focusing on any areas that didn’t meet the prompt requirements.

Confirm that the rewrite aligns closely with the task’s objectives and expectations.