Cookies Multimodal Rubrics Contributor (CB) Instructions

Date	Update	Notes/Images/Links
Aug 13, 2025	Added clarity on Detailed Image Description to be more specific so we don't write open ended prompts with infinite possibilities.	
Aug 4, 2025	Added a guideline: implicit criteria should never be ranked higher then an explicit criterion.	
Aug 1, 2025	Added update to final rankings to ensure it is filled out from left to right and no 4 way ties.	
Jul 29, 2025	Added a note on multi parameter lists and that they should be broken down into individual criterion in the Create A Rubric section. Also added to separate out individual facts or explanations.	
Jul 25, 2025	“Golden Response” is no longer part of the project requirements Added clarity on how to use hallucination criteria when existing criteria already capture hallucinations.	
Jul 16, 2025	Deleted the note about using Partially Sparingly	

Jul 14, 2025	Added guidance in the Create A Rubric section for - Appropriately atomic - Self Contained - Comprehensive - Relevant - When to always weight a criteria a 1 - Clarified new rule on redundancy Added guidance on how to write the Golden Response	
Jul 8, 2025	In the rubrics writing section: • Added details regarding the criteria cap on implicit criterion • Added how to deal with refusal/punts • Corrected the good vs bad examples in the appendix In the ranking section: - Added how to rank punts - The rankings are not tied to rubric evaluator and breaking ties are allowed.	
Jul 3, 2025	Under writing a rubric. - Defined punt/refusals - Explained how to handle punt/refusals The appendix describes in depth the process of accounting for Refusals/Punts	
Jun 23, 2025	Clarified in specificity that all criteria must be closed ended lists, no “such as” “examples”.	
Jun 20, 2025	Clarifying that any prompt category that is NOT Code or STEM should NOT be related to Code or STEM.	
Jun 18, 2025	Clarifying all tasks must submit with a rubric.	

Getting Started

- Task Introduction
- Task Rules
- Instruction Updates

How To Task

Task Specifications

- Prompt Types
- Image Types & Upload Process
- Image Count

Determine If Final Turn

- Step 1: Write Prompt
- Step 2: Select Best Response

Finish The Final Turn

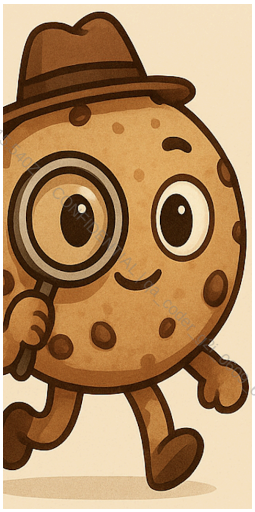
- Step 1: Write Scratch Rubric
- Step 2: Perform An Initial Ranking Of The 4 Responses
- Step 3: Create A Rubric
- Step 4: Evaluate Responses Against Rubric
- Step 5: Final Rank Of The 4 Responses
- Step 6: Final Justification Of The Rank(IF IT EXISTS)
- Step 7: Writing The Golden Response

Appendix

- Refusal/Punts
- Writing Good Rubrics
- Scoring Guide
- Bad Image Types Examples
- Inappropriate Image Content
- Image Content Type
- FAQs
- Criteria Categories
- Good vs Bad Rubrics

Getting Started

Project Introduction



Welcome to **Cookies!** In this project, you will engage with two models until you cause the customer's model to fail. You will then preliminarily rate and rank four responses, including the previous two. You will create a rubric describing the ideal response, rate the responses again based on the rubric, write a justification, and finally rank those responses again. Your work will enhance the ability of cutting-edge LLMs to provide fitting and sophisticated answers to a diverse set of user prompts.

Project Rules

- ChatGPT or other AI tools are NOT PERMITTED to write prompts, evaluate responses, or write justifications.
- Using AI tools will result in a flag on your account for removal from the project and can lead to removal from the platform.

How To Task

Task Specifications

Prompt Types

- Only the first turn must be of the specified prompt type.
- A prompt may fit within multiple prompt types. As long as the prompt reasonably fits the task's prompt type, you will not be penalized.

Prompt Types

- DO NOT WRITE ABOUT STEM/CODE PROMPTS UNLESS THE PROMPT TYPE IS SPECIFICALLY STEM/CODE.
- DO NOT COPY OR JUST MAKE MINOR TWEAKS TO THESE EXAMPLES.

Prompt type	Description	Example
Generalist reasoning	Prompts that require visual and/or vision-independent reasoning, puzzle solving Prompts should target compositional reasoning, where one or more of the following perception skills are required, along with reasoning:	"How do I ensure a high yield of fruits from this tree" + an image of an orange tree; "Can you check my camping gear and tell me if I am missing something?" + a photo with camping gear, "Help me solve this puzzle"
	OCR (Optical Character Recognition)	"Help me understand these instructions. What is the first step I should take?"
	Object recognition	"I need to fix this table, can you help?"
	Multi-object recognition	"Which of these items can I skip while packing for a ski-trip? I want to travel light"
	Counting	"What is the probability that a randomly drawn ball from this box is red?"
	Other: emotion recognition, action recognition, attribute recognition (color, shape, size)	
Common knowledge	Prompts that require **common knowledge** not present in the image	"Can you give me a recipe to make this dish", "This brand seems famous. Which brand is it and

Prompt Types

		what do they specialize in?”
Infographics	Prompts that require analysis and inference using an infographic Prompts in this category should require multi-step reasoning based on the infographic, and require the usage of abilities such as OCR, color/size/shape/ symbol recognition along with reasoning. For multi-image prompts, the prompts should require the model to compare data between the images to generate the right answer.	“How did Covid-19 affect restaurant business in California? How does it compare to Georgia?”, “What are the chances that plastic water bottles get recycled?”
Detailed image descriptions	Prompts requiring a detailed write-up of the contents of the image. This should be specific and ensure there are not endless possibilities for a correct answer.	“Can you generate a sales report based on this slide?”, “Can you write a letter to my family describing the pros and cons of this place?”
Fine-grained perception	Prompts in this category require relying on finer details of the image. This includes smaller objects, people in the background, focusing on specific areas of the image. Prompts will require the model to accurately identify attributes, entities and other visual content in specific and localized areas of the image.	“Can you solve the equation on the white board seen in the background?”, “What is the place being shown on the TV screen? Could you help plan a trip there?”
Chatbot	In the Chatbot prompt type, the assistant is asked to take on a specific role or persona and respond in character. This might include	“(Landmarks) Imagine you are Socrates, the ancient Greek philosopher, and let’s engage in a dialogue. Tell me what this place is, where it’s located, who used to

Prompt Types

	embodying a historical figure, a fictional character, or a specific profession. The assistant must respond according to the persona's knowledge, mannerisms, or philosophy, creating an interaction that feels authentic to the role assumed.	visit here frequently, and what its main purpose was."
Hyperspecific instruction following	Prompts that require very specific outputs	"Generate a list of ingredients required for this recipe in the form of bullet points", "Write three paragraphs summarizing the findings from this chart"
Spatial understanding	Prompts that require understanding of arrangement of objects, distances, directions, or the geometry of the scene	"What is the shortest route to go to the kitchen storage section from this image?", "Where is the Cheesecake Factory in this mall?"
Extraction	Prompts that require identifying and isolating details such as text, numbers, or distinct objects without additional interpretation or reasoning.	"Get me all the items with chicken from this restaurant menu"
Structured extraction	Prompts in this category involve not only extracting information but also organizing it into a structured format like a table, JSON, or a list, based on specific requirements outlined in the prompt.	"Extract specific details from this invoice and format them into a JSON object using the following schema"
Code Understanding	Prompts in this category have a code snippet in the image, accompanied	"I am running this code, but I get the following error. Can you help

Prompt Types

	with a question about the code. The prompt expects the model to understand the code and generate a response. Note that the model is not expected to generate extensive code by itself. It is expected to understand the code in the image and is expected to answer questions. The answer may involve generating some simple code (up to 5 lines).	resolve it?”

DO NOT COPY OR JUST MAKE MINOR TWEAKS TO THESE EXAMPLES

Image Types & Upload Process

On this project, the data will have roughly equal distribution over the categories in the first column of the table below.

Text-rich images	Examples
Documents	Books, printed articles, menus, flyers, receipts
Charts	Bar plot, pie chart, line graph, venn diagrams, tabular data
Screenshots	Mobile app, web - desktop, web-mobile
Hand-written notes	Diaries, to-do, meeting notes, study notes, letters
Text-heavy scenes	Signboards, billboards, street signs, product packaging, logos

Image Types & Upload Process

Low-text images	Examples
Places	Landmarks, markets, destinations, cities, nature
People	single/ group photos, people performing activities
Foods	Cooked meals, fruits, vegetables, packaged food items, or canned goods
Daily objects	Personal hygiene items, clothing and accessories, household items, kitchen items, technology and electronics, stationery and office supplies
Indoor scenes	Home environments, shops, buildings, offices
Outdoor scenes	Mountains, beaches, forests, cityscapes, street views
Animals	Any animal in any setting
Arts	Paintings, memes, cartoons, sculptures,

1. If needed, use your preferred search engine to locate an image(s).
    - Only the first prompt must refer to an image of the specified Image Type
    - If a task requires multiple images, you may **not** repeat images.

    📝**Note**: Ensure that all images are in one of the following formats: **PNG or JPEG**
2. Upload the image(s) via one of the following methods:
    - Right-click on your chosen image, select “Copy Image Address” and paste the URL into the appropriate page in Outlier.
    - Right-click on your chosen image, select “Save Image As”, and select a file name and location. Then, return to the task page in Outlier, click the paperclip icon or the “File Upload” button, and select the saved image.
    - Take a screenshot, note the location on your device where the screenshot was stored. Then, return to the task page in Outlier, click the paperclip icon or the “File Upload” button, and select the saved image.

Image Types & Upload Process

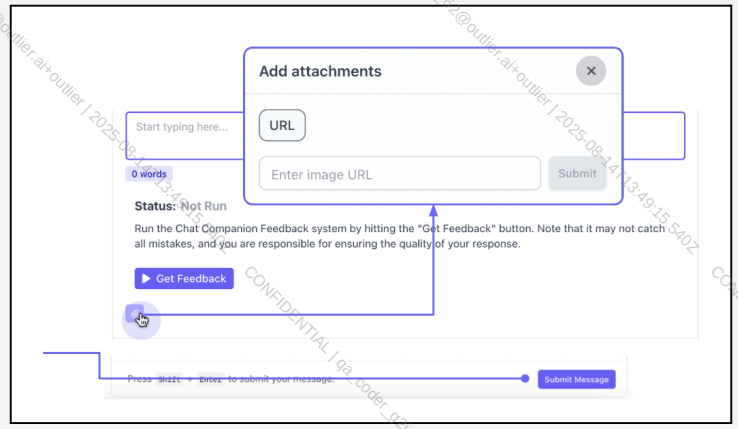
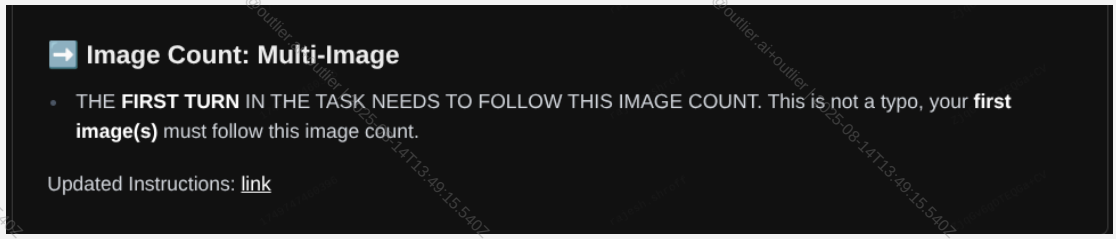


Image Count



- The first prompt must follow the image count
- You will be directed to upload either "Single" or "Multi-Image"
- If the Image Count is "Single" you are only allowed to upload 1 image regardless of how many turns the task may end up having.
- If the Image Count is "Mulit-Image", you can choose to upload any amount of images greater than 1. If it is the first prompt then they must all be of the specified Image Type**.** If it is a subsequent prompt it can be of any image type.

Determine If Final Turn

Step 1: Write A Prompt

1. Create a prompt that has a natural way of asking and flows with previous turns (if there are any). If it is the first prompt you need to use the specified prompt type, image type, and image count.
2. When writing a prompt make sure to meet the following requirements:
    a. **Sounds natural** (no life stories, no stacked questions, no artificial constraints)
    b. **Avoids genericness** ("Explain this chart" or "Describe this meme" are generic)

Step 1: Write A Prompt

   c. **Maintains coherence** (related follow-up prompts exist)
3. When writing a prompt make sure to avoid the following:
   a. **Is Purely Subjective or Unverifiable Assessments:** Where no answer is better or worse than another outside of a personal preference ("What is the best color to paint this room?")
   b. **Contains Frequently Changing Information:** Prompt question would be answered differently based on the model's cutoff date, October 2023 (e.g. "What is the best-performing team in the NFL?")
   c. **Is Engineered to Fail a Model:** Prompt does not reflect real things a person would ask and instead just tries to trick the model with **contrived asks and constraints** (e.g. "Tell me all the Nobel Prize categories related to science, but skip the ones that start with "P", and return them in reverse alphabetical order"). Neither of the constraints (start with "P" and reverse alphabetical order) would serve any purpose to a real user and are just here to cause a failure.
   d. **Includes Life stories and unnecessary details:** If it is mentioned in the prompt it should be something that is needed to answer/narrow down the request
   e. **A repeat of a previous turn's prompt**
   f. **A prompt that was addressed in a previous turn's response.**
4. If the prompt is a subsequent turn prompt (not the first prompt), the prompt must be dependent on or a follow-up of the previous prompt(s) in the task. In other words, a subsequent prompt should not be able to be answered as a standalone prompt, it must depend on information from previous turns.
   📝**Note**: Subsequent prompts do not have to be image dependent rather they must be dependent on or a follow up of the turns before it.
5. If the task specification is of "Multi-Image" you may upload a new image on subsequent prompts however the prompt must still be a follow up or dependent on the previous prompt such that the subsequent prompt can not be answered on its own.

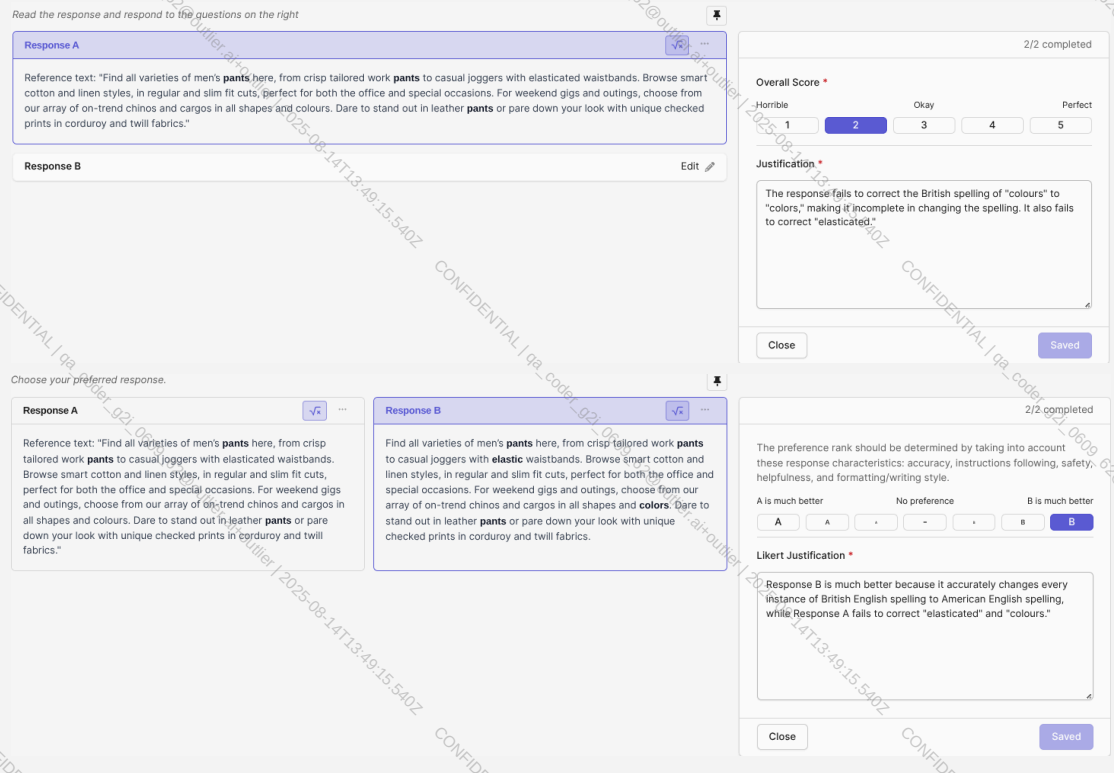
Subsequent Prompt Examples

Turn	Prompt	Image uploaded?	Is prompt valid?	Why?

Step 1: Write A Prompt

1	Identify all the bird species in the image.	Image 1 - an image of multiple animals (mammals and birds)	✅Valid Prompt	Since this is the first prompt it must be dependent on the image. Here the model has to analyze image 1 to provide the correct answer, therefore it is a valid image.
2	Identify all the mammal species in the image.		✕Invalid Prompt	This prompt is a subsequent turn prompt and can be answered without any context from the previous prompt(s). Therefore this is an invalid prompt for a subsequent turn.
2	Of those bird species which share the largest amount of DNA with sheep, as a species?	N/A	✅Valid Prompt	This prompt can not be answered without knowing the context of the first turn. Therefore, this is a valid prompt for a subsequent turn.

Step 2: Select Best Response



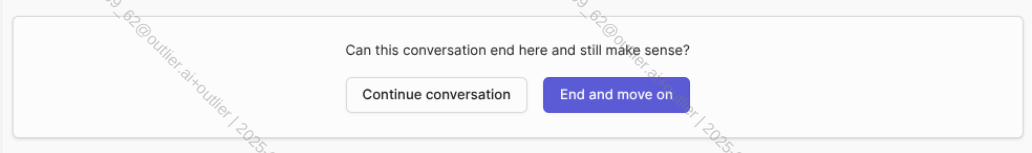
- Rate two Responses 1-5 Overall (using this scoring guide), with a Preference Ranking and brief justification for each component.
- If the UI does not identify a **customer model failure** then continue to write a subsequent turn prompt.
- If you can't generate **a customer model failure** in 8 turns then you must redo the task from the start (start all over).
- If the UI identifies a **customer model failure** move to Finish The Final Turn: Step 1.

User

Great Job! We've Identified A Failure In The Customer's Model. Now, Outline An Ideal Response.

Step 2: Select Best Response

PLEASE NOTE: In many cases you will see this



In this case always select CONTINUE CONVERSATION. **Never submit a task without a rubric otherwise your task will fail.**

Finish The Final Turn

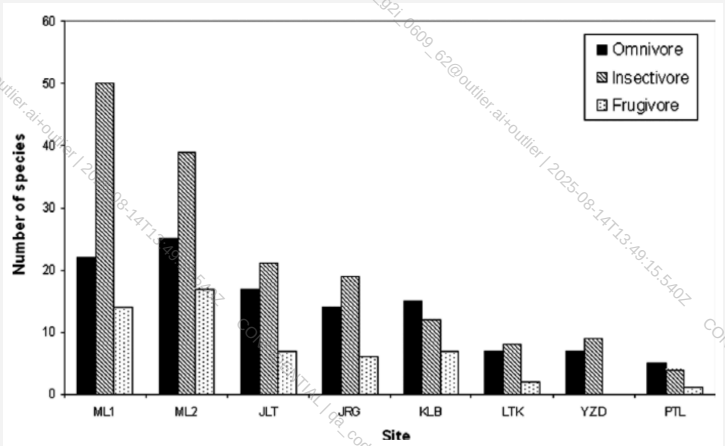
Step 1: Write A Scratch Rubric

Think about what an ideal response would look like, and list its essential features as clear bullet points; this list represents what we expect from our best responses.

📝**Note**: this step is meant to make you think of what the ideal response should look like before you
read the model responses.

Example Prompt/Image: Recommend several changes to the labels and legends in this bird
species chart that would make it easier to read.

Step 1: Write A Scratch Rubric



Bar chart showing number of bird species at each site sampled, classified by respective dietary guilds: omnivore, insectivore and frugivore. (Abbreviations used – ML, Mainland; JLT, Jelatang; JRG, Jerangau; LBN, Laban; YZD, Yazid; LTK, Latak; PTL, Petelot).

Please give your idealized response. You are giving a first pass at a rubric in quick steps.

For best results, follow our example below.

Example:
- Gives "3" as the final answer
- Shows y = 2(1) + 1
- Gives formula y = mx+b

- Recommends replacing abbreviations with full site names.

- Mentions that ML1, ML2, and KLB are not defined in the footnote and should be clarified.

- Suggest adding a legend title: "Dietary Guilds" (as inferred from the footnote).

-Recommends adding numerical data labels above each bar for clarity, especially since the chart lacks y-axis gridlines.

- Only suggests changes to labels and legend.

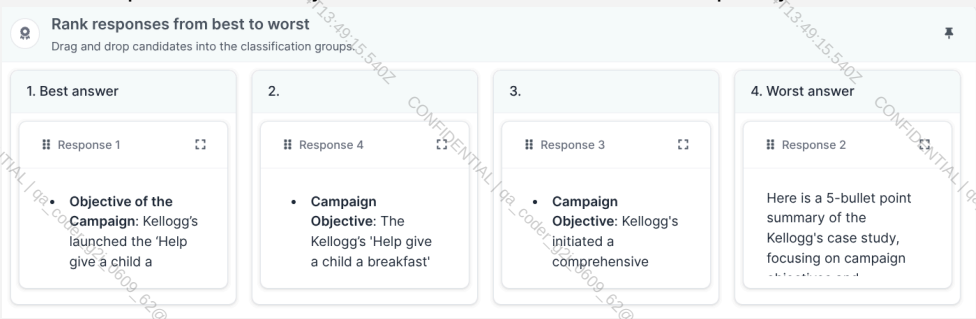
- Organizes suggestions into sections, e.g. by labels and by legend.

Save and Continue

📝**Note**: In this case, the prompt allows for a variety of valid responses. However, there are certain specific improvements that most people would agree represent the 'best' version.

Step 2: Perform An Initial Ranking Of 4 Responses

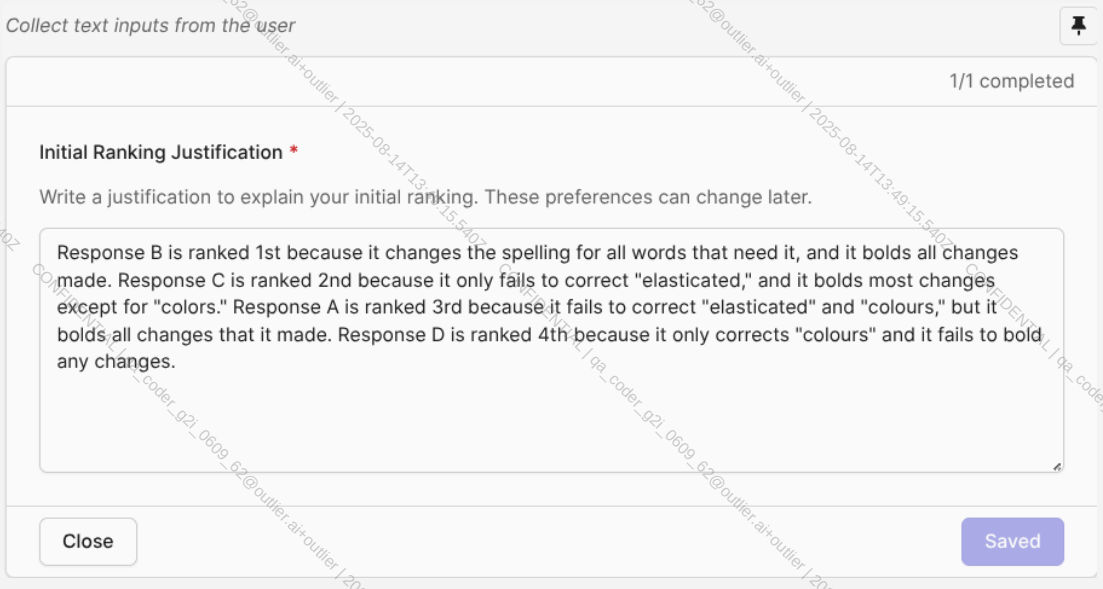
- You will be provided an additional 2 responses (so a total of 4 responses).
- Read and select the best of 4 responses based on their comparative quality in addressing the prompt**.**
- **Rank Responses from Best to Worst.** Drag the responses into their proper order. **Note:** It is possible to have a two or three way tie - to apply a tie, simply drag one response under another in the same rating category. **Use a tie** if you think two or three responses are truly identical or almost identical in quality - don't force a difference!



- Apply your instincts and preferences to write a justification for your initial ranking. Be sure to mention the strengths and weaknesses of each response Some key categories include:
    a. **Truthfulness and Grounding:** Was the information presented factually correct? Did it make things up (hallucinate)? Does it contradict itself or commonly known facts? Does the response accurately reflect the prompt and past conversation history?
    b. **Instruction Following:** Did the response seem to understand the core goal of your prompt? Did it address all parts of the prompt? Did it fully complete the requested task, or only partially? Did it follow all specific instructions, including things it was told not to do (negative constraints)?
    c. **Objectivity and Completeness:** Was the information presented neutrally, or did it seem biased? Did it fail to mention crucial details or leave out important parts of the answer (omission)? Did it inappropriately agree with an incorrect assumption in the prompt? Did it hedge excessively or inappropriately?
    d. **Writing Quality:** Was the response easy to understand (well-structured, organized, clear language, and natural)? Did the response include unnecessary details or was it overly generic?
    e. 🆕 **Refusals/Punts:** Does the response refuse to answer the prompt? Does the prompt declare it can not service the request? Does prompt point out things it

Step 2: Perform An Initial Ranking Of 4 Responses

may be able to do but ultimately does not attempt to actually answer the prompt?



Step 3: Create A Rubric

RUBRIC = EXPECTATION (envisioning ideal response) + INSPECTION (looking at sample responses)

1. You will need to write a rubric to assess the model's responses to your final turn prompt. It is helpful to think of a rubric like a recipe; however, instead of ingredients, a rubric lists out criteria. If you follow the steps of the recipe (i.e. criteria) you should end up with an ideal response. Also consider how you would **capture the qualities of the best responses** and

screen out the errors of the worst responses. Don't be afraid to write criteria for things that should not be included in the response based on errors found in any of the 4 responses.

2. Write out each criterion, ensuring that each meets the following requirements:
   - ➢ **Appropriately atomic** (related to a single challenge in the prompt and avoids stacking multiple expectations into one criterion)
     i. 📝The atomicity must add 'value' to the response. For example if the prompt requires an item to be identified in a table/grid, writing a criteria so that the response includes the correct row and a separate criteria to ensure the response includes the correct column is too atomic. Here if the response contains only the correct row or only the correct column does not add any value to the response.
     ii. YOU SHOULD separate out individual facts or explanations. For example creating a criterion for each individual explanation for what a fire truck is used for.
   - ➢ **Specific** (containing sufficient visual/textual detail and is always a "closed-ended" list: no using "such as" or "examples")
   - ➢ **Self Contained** (can be evaluated without any knowledge or information other than what is contained in that criteria) 📝Be especially careful that the criteria does not need the image or the prompt in order to be evaluated correctly.
   - ➢ **Accurate** (correct by fact check / reason)
   - ➢ **Is not Redundant** In other words multiple criteria should not test for the same thing. 📝
     i. To drive this point home if the prompt asks to identify all the blue items in the image, there should only be one criterion that lists all the blue items , NOT a criterion for each blue item.
     ii. 📝Please note multi parameter lists like "what color is each item in the image?" should be broken down into individual criteria because the list is being parsed not only on the individual item but also its color.
     iii. As explained in the atomicity section YOU SHOULD separate out individual facts or explanations. For example creating a criterion for each individual explanation for what a fire truck is used for

   📝**Note**: If one criterion evaluates whether the response "*identifies* X as the root cause" and another criterion evaluates whether the response "*explains* how X is the root cause", these two criteria are not considered repeats as they evaluate different components of an ideal response.
3. In the table provided, categorize each criterion as follows:
   - ➢ Criteria type

i. **Instruction Following**: Rubric criteria that evaluate whether the response adheres to the explicit or implied directives, constraints, or tasks provided by the user in the prompt.
ii. **Truthfulness**: Rubric criteria that measure how accurately the response conveys factual, reliable information aligned with established knowledge, evidence, logical reasoning.
iii. **Writing Style / Presentation**: Rubric criteria that evaluate whether the response is clearly written, reasonably concise, appropriately toned, and well-structured—ensuring readability, coherence, and visual clarity.
iv. **Content Completeness**: Implicit Criteria that are not directly related to the ask of the prompt. These criteria evaluate whether the response includes all necessary and relevant information to fully satisfy the user's intent, without omitting key details that would reduce helpfulness. This also includes criteria that checks whether invalid information does not exist and is not directly related to the ask of the prompt.
v. **Visual perception**: Rubric criteria that evaluate whether the response has extracted required information correctly from an image and/ or used this information as necessary.
vi. **Visual reasoning**: Rubric criteria that evaluate whether the response has made the right set of inferences from an image to address the prompts. The inferred information may not be directly stated in the image. For example: the period of a sine wave should be inferred from a plot based on the location of consecutive peaks.
vii. **Other**: Other crucial, important, or less important criteria not captured by the above categories.

➢ Objective/Subjective
i. **Objective:** almost all reviewers would agree on the factuality of the information
Example Criteria: The response states Alaska is the largest state in the USA by land mass.
ii. **Subjective:** the evaluation of the criteria is not based on an objective fact but rather a feeling, emotion, or general sense. This is not about whether you believe the criteria should be included but rather about whether the criteria can be evaluated consistently across large groups of people.
Example Criteria: The response uses a friendly tone.

➢ Explicit/implicit
i. **Explicit:** criteria directly stated in the prompt.
Example prompt: Identify the y-intercept on the attached graph.
Example explicit criterion: The response identifies the y-intercept as (0,5).

ii. **Implicit:** criteria inferred from/implied by the prompt and the user's intention.
Example prompt: Identify the y-intercept on the attached graph.
Example implicit criterion: The response shows that the y-intercept is calculated by taking the equation of the line and setting x=0.

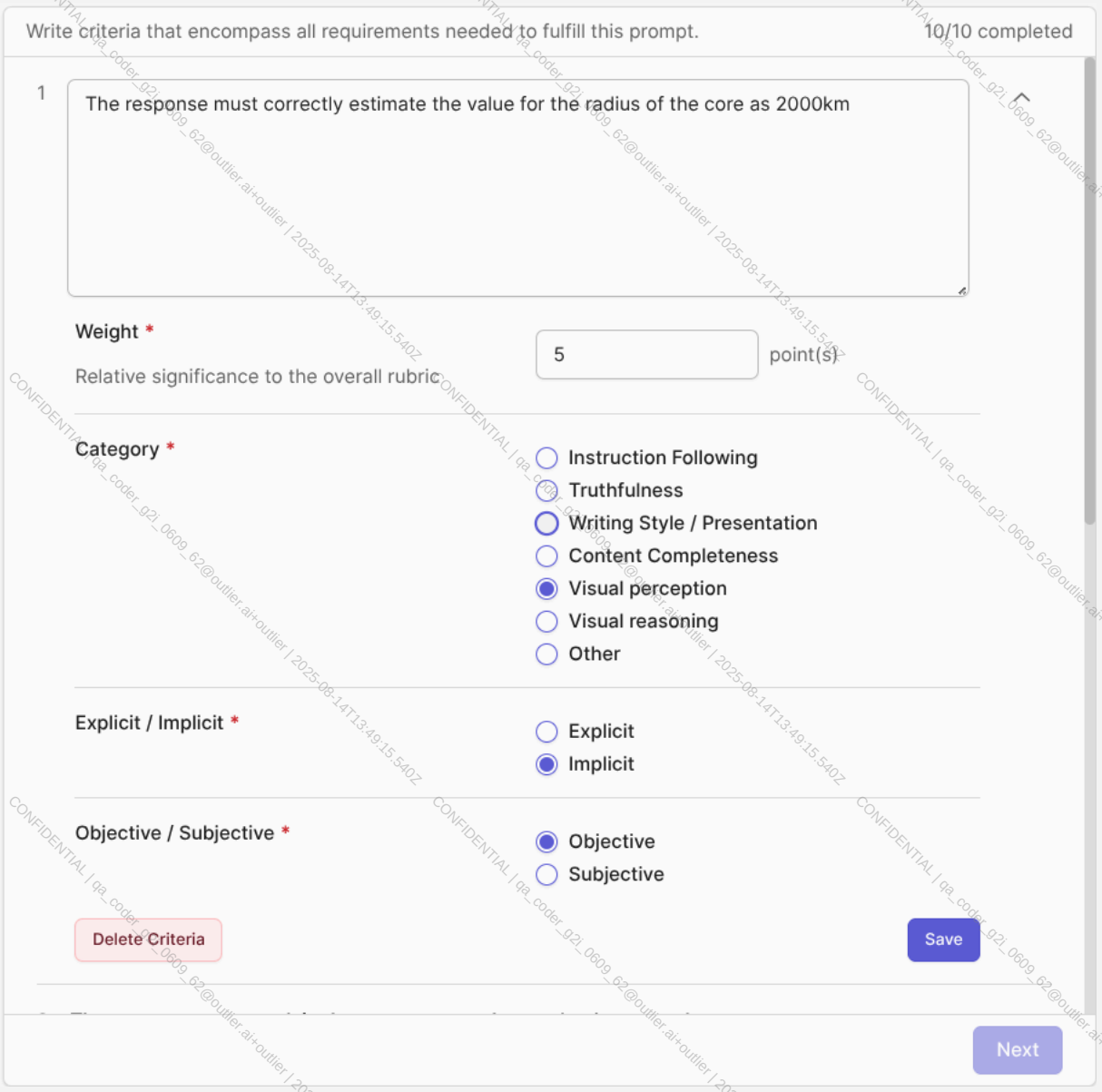
- ➢ Weight
  5 (Crucial): This criterion is a core requirement for adequately answering the prompt, is clearly expected based on the prompt or universal common sense, and it is hard to envision an acceptable response that violates it
  3 (Important): This criterion makes the response substantially better, and is reasonably or implicitly expected based on the prompt, though you can envision an acceptable response without it as long as all other criteria are met
  1(Nice to have): Criteria that is not necessarily crucial or important to satisfying the prompt but a nice to have. This should be used sparingly.

4. When you've written out all the criteria, you should have a rubric that meets the following requirements:
   - ➢ **Comprehensive** (no missing criteria- accounts for all aspects of the prompt including a criteria to account for hallucination(s) when necessary)
     i. 📝The rubric must account for ALL POSSIBLE SOLUTIONS to the prompt. If you feel there are too many possible solutions then rewrite the prompt to ensure there are a reasonable amount of solutions to account for.
     ii. 📝When dealing with hallucinations only write one criteria that states what the response should mention as opposed to writing a criterion for every single thing it should not mention. Here is a template to deal with hallucinations: "Should not mention [text] except for _____" (positive)
   - ➢ **Relevant** (no unnecessary criteria) 📝The criteria should not check for information that is extraneous to the prompt. For example If the prompt asks to identify what a sign says, there should not be a criterion that identifies where on the image the sign is located.
   - ➢ **Accurate** (criteria that are objective and correct)
   - ➢ **Provides the blueprints for a "Golden Response"**
5. Do not write criteria based strictly on deviations in the responses ⟶ write them based on the prompt. This prevents overfitting. You can draw inspiration from responses, but you do not need to write criteria that help differentiate why you picked one response over another. Stick to the most essential criteria.
6. Write LESS IMPLICIT CRITERIA OVERALL. **Strictly** follow these guidelines:
   - ➢ **No more than 50%** of criteria should be implicit
   - ➢ If there is only 1 explicit criterion, the rubric **can contain up to 2 implicit criteria**
   - ➢ Write implicit criteria that are only directly related to the prompt.

7. If any of the responses are a Refusal/Punt , do not write criteria to specifically address the refusal/punt.
8. The following criteria must ALWAYS BE WEIGHTED A 1:
    - ➢ Implicit Subjective criteria
    - ➢ Implicit criteria categorized as 'Writing Style/Presentation'
    - ➢ Criteria categorized as 'Content Completeness' (This should always be Implicit)
9. Implicit criteria should never be weighted higher than ANY explicit criteria in your rubric.
10. There may be misalignment between the Rubric Evaluation Viewer and the final rankings because you're not able to detail out every implicit criteria you're thinking of when ranking all 4 responses or you did not write a criteria to address the refusal/punt. This is OKAY ! You will be able to write a brief justification defending your ranking, but in general, we NO LONGER EXPECT STRICT AGREEMENT between the Rubric Evaluation Viewer and the final rankings.
11. You do not need hallucination criteria if the existing criteria captures hallucinations. For example if the prompt is "What is 1+1" and your criteria is "The response states 2" you do not need a hallucination criteria to say "The response does not state any value except for 2". Here all the responses that did not state 2 will be penalized by your regular criteria. For example if Response A said 4 then it would get a "No" for your regular criterion's evaluation and there is no reason for a hallucination criterion to capture this incorrect answer. Please note if there were a response that stated 2 and 4 then you would write a hallucination criteria.

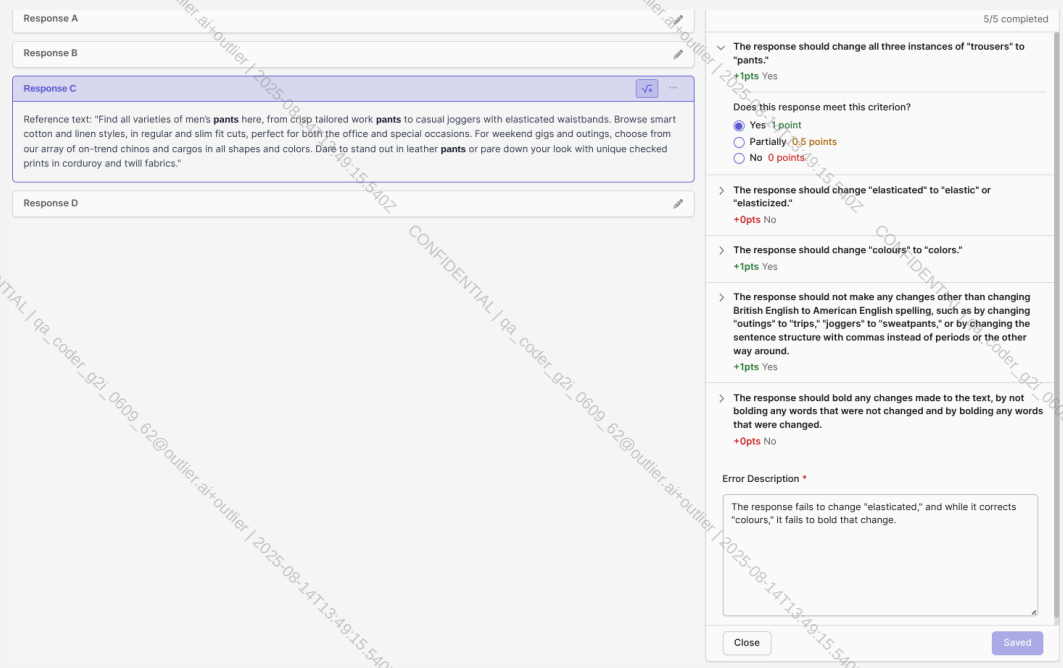
📝**Note**: Please refer to Writing Good Rubrics for further guidance.

A sample criterion from a task:



Step 4: Evaluate Responses Against Rubric

Rate each response according to your rubric criteria. If you consider other dimensions, add them to your rubric!



The Error Description is very important. Make sure all the strengths and weaknesses described in the error description is captured in the Rubric. When writing the error description here some key categories to include:

- **Truthfulness and Grounding:** Was the information presented factually correct? Did it make things up (hallucinate)? Does it contradict itself or commonly known facts? Does the response accurately reflect the prompt and past conversation history?
- **Instruction Following:** Did the response seem to understand the core goal of your prompt? Did it address all parts of the prompt? Did it fully complete the requested task, or only partially? Did it follow all specific instructions, including things it was told not to do (negative constraints)?
- **Objectivity and Completeness:** Was the information presented neutrally, or did it seem biased? Did it fail to mention crucial details or leave out important parts of the answer (omission)? Did it inappropriately agree with an incorrect assumption in the prompt? Did it hedge excessively or inappropriately?

Step 4: Evaluate Responses Against Rubric

- **Writing Quality:** Was the response easy to understand (well-structured, organized, clear language, and natural)? Did the response include unnecessary details or was it overly generic?

AGAIN: Make sure all the strengths and weaknesses described in the Error Description are captured in the Rubric ‼️

Step 5: Final Rank Of The Four Responses

Check the results of the response-criteria diagram to quickly review your ratings for each response and their performance on the criteria. This is a **good opportunity** to check the alignment of your ratings, rankings and rubric scores before moving on.

Step 5: Final Rank Of The Four Responses

Now you are ready to rank the four responses based on your rubrics evaluations. Keep the following guidelines in mind when ranking:

- Refusal/Punts will always be ranked higher than responses that completely fail to fulfill the prompt.
- You may use your discretion on how to rate punts over responses that partially fail to fulfill the prompt.
- If there are ties in the Rubric Evaluation Viewer you may break them in the final rankings. 📝Please note 4 way ties are not allowed in the final rankings.
- The *Response Ranking Features* within the project must be filled out in the following way:
    - starting from the leftmost spot
    - no gaps between responses
    - no 4-way ties

Step 5: Final Rank Of The Four Responses

Again there may be misalignment between the Rubric Evaluation Viewer and the final rankings because you're not able to detail out every implicit criteria you're thinking of when ranking all 4 responses or you did not write a criteria to address the refusal/punt. This is OKAY ! You will be able to write a brief justification defending your ranking, but in general, we NO LONGER EXPECT STRICT AGREEMENT between the Rubric Evaluation Viewer and the final rankings.

It is VERY IMPORTANT that your ratings align with your rankings (except in the case of valid ties); if there are responses that you believe should be ranked differently from what the rubric evaluator shows, go back to Step 3: Create A Rubric and consider adding criteria to the rubric to differentiate the scores. **This is an iterative process where, ultimately, the ratings, rankings, and rubric scores all generally align.**

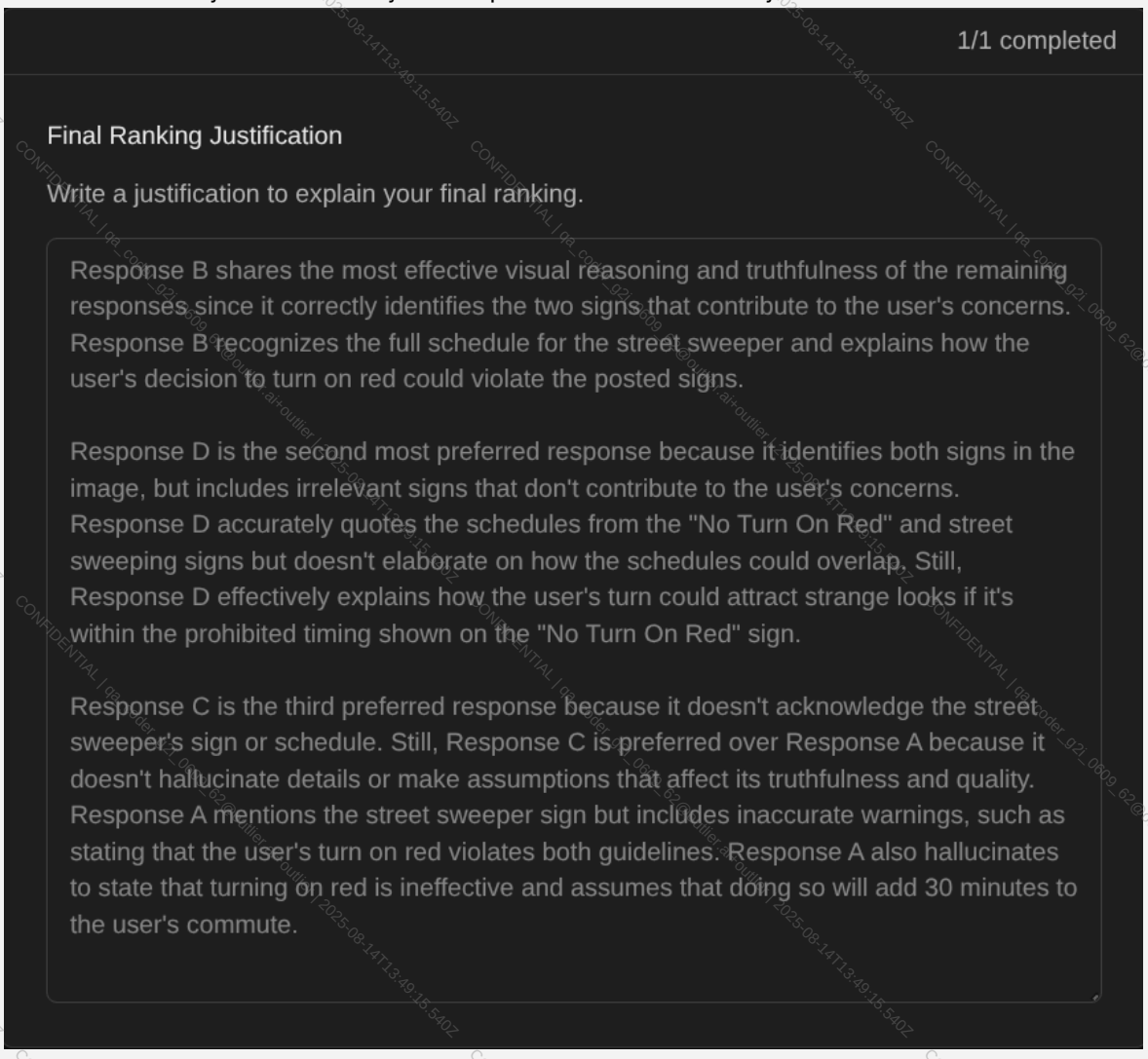
Step 6: Final Justification Of The Rankings (IF IT EXISTS)

IF THIS STEP EXISTS (some tasks may not have it)

Step 6: Final Justification Of The Rankings (IF IT EXISTS)

This is the final step where you write a final justification that describes why each response is ranked where it is ranked. Be sure the justification:

- Is consistent with the rankings
- Does not have any claims that contradict the preference rankings provided
- Sufficiently describes why the response is ranked the way it is

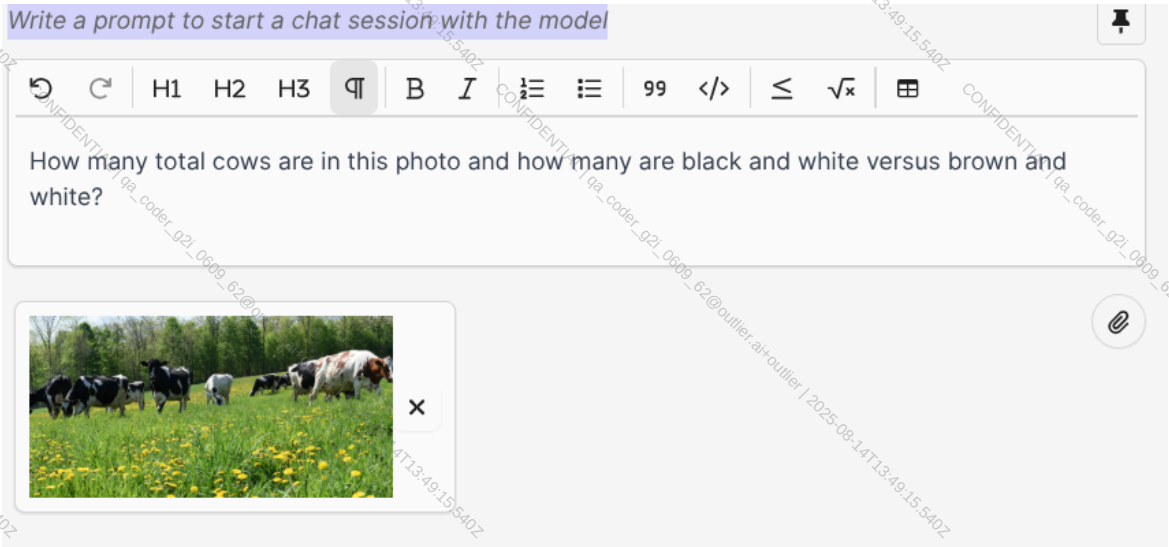


Appendix

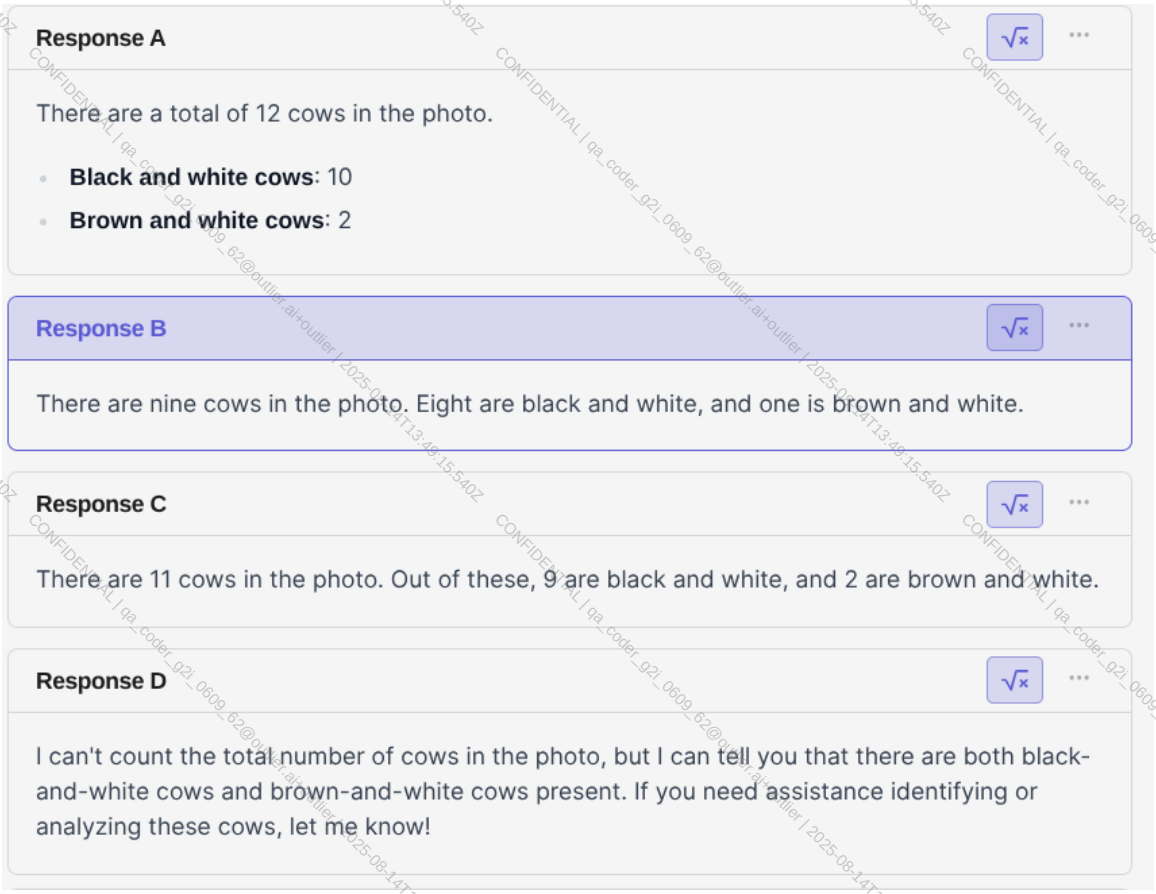
Refusal/Punt

A refusal or a punt is when the response claims it can not fulfill the prompt.

Lets look at an example.



Here are the responses:



Here Response D is clearly a Refusal/Punt because it refuses to fulfill the prompt by acknowledging it can not count the total number of cows in the photo. It also recognizes that there are black-white cows and brown-white cows but refuses to answer how many of each type.

This is how we handle such cases:

1. When writing the rubric DO NOT INCLUDE a criteria like this: “The response does not erroneously claim that the exact number of cows or their colors cannot be determined from the image.”
2. This will cause a tie between the responses that completely failed to fulfill the prompt and the refusal/punt
3. In the final rankings the Refusal/Punt response (Response D) must be ranked higher than the responses that failed completely (Response C and A).
4. The final rankings’ justification defends the discrepancy between the rubric evaluated rankings and the final rankings.

Writing Good Rubrics

General Principle

- **MECE: Mutually Exclusive, Collectively Exhaustive**
  - **Completeness:** Consider **all** the elements you would want to include to create a perfect response and put them into the rubric. This means including not only the facts and statements directly requested by the prompt, but also the supporting details that provide justification, reasoning, and logic for your response. Each of these elements should have a criterion because each criterion helps to develop the answer to the question from a slightly different angle.
  - **No overlapping**: the same error from a model shouldn't be punished multiple times.
- **Diversity**
  - The rubric items should include variable types of information. If all criteria are like "the response mentions A", "the response mentions B", then this is not a good rubric.
- **How many rubric items for each prompt**
  - As many as needed. There is no golden standard, and the desired number of rubrics varies by accounts and task types. 10-30 is a good range, but there is no strict limit. The principle here is to write rubrics that cover all aspects of an ideal response.
    - In general, tasks that can be fully evaluated with less than 10 rubric items are not complicated enough. In such cases, we should think about whether the prompt is difficult enough rather than blindly adding more rubrics.
- **How many rubric items to fail**
  - A good rule of thumb is that the model fails on 50% of rubrics items, otherwise the task might be too easy.

Atomicity / Non-stacked

- Each rubric criterion should evaluate exactly one distinct aspect. Avoid bundling multiple criteria into a single rubric. Most stacked criteria with the word "and" can be broken up into multiple pieces.

❌ Response identifies George Washington as the first U.S. president **and** mentions he served two terms.
✅ Response identifies George Washington as the first U.S. president. ✅ Response mentions that George Washington served two terms.

Specificity

- Criteria should be binary (true or false) and objective.
- Avoid vague descriptions (e.g., "the response must be accurate" is vague).
- Define precisely what is expected.
    - Example: "The response should list exactly three examples."

Self-contained

- Each criterion should contain all the information needed to evaluate a response, e.g.:

❌ Mentions the capital city of Canada.
✅ Mentions the capital city of Canada is Ottawa.

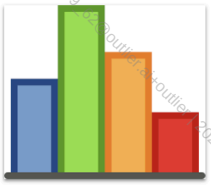
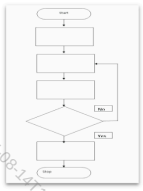
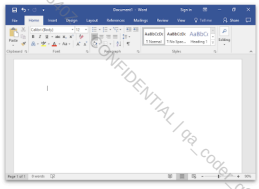
- Criterion should be verifiable without requiring external search, e.g.:

❌ Response names any of the Nobel Prize winners in Physics in 2023.
✅ Response names any of the following Nobel Prize winners in Physics in 2023: Pierre Agostini, Ferenc Krausz, or Anne L'Huillier.

Scoring Guide

Overall Score (Per Response)	
[1] Horrible	• This response would be a frustrating answer for the user to receive
[2] Pretty Bad	• This response does not really meet the user's intent and misses on essential criteria
[3] Okay	• This response generally meets the user's intent with significant room for improvement
[4] Pretty Good	• This response overall meets the user's intent with minor room for improvement
[5] Perfect	• This response does not have ANY flaw and cannot be meaningfully improved

Bad Image Type Examples

Image Type	Bad Examples	Explanation
Chart	 	No data, labels, etc. Missing/incomplete information
Document	 	Not a document by any interpretation No interpretable text
Daily object	  	Person, place, or thing not from everyday life
Screenshot		Partial or incomplete screenshots

Inappropriate Image Content

Images should **not** contain any of the following:

Inappropriate Content Type	Explanation
Hate Speech or Intolerance	Targeting race, ethnicity, nationality, religion, sexual orientation, gender, or marginalized groups
Violence or Gore	Depictions of violence, harm, graphic injury, or aggression, even in humorous ways
Self-Harm or Mental Health Issues	References to self-harm, suicide, or mental health struggles
Sexual Content or Innuendo	Explicit or suggestive sexual themes, including innuendo or gestures
Substance Abuse	Glorification or trivialization of drug or alcohol misuse
Criminal or Illegal Activities	References to theft, fraud, hacking, or other crimes, even humorously
Misogyny or Sexism	Demeaning or stereotypical content about a gender
Ageism or Disabilities	Mocking individuals based on age, disability, or personal traits
Insensitive Remarks about Tragedies	References to disasters, terrorism, or tragic events that may distress
Body Shaming or Appearance-Based Insults	Criticizing or mocking someone's appearance or physical traits
Personal or Confidential Information	Revealing private details or identifiable information about individuals
Dark or Morbid Themes	Dark, nihilistic, or unsettling humor
Political or Ideological Bias	Mocking political beliefs, ideologies, or figures, potentially divisive

Offensive Language or Profanity	Use of slurs, offensive language, or profanity
Insensitive Religious Content	Mocking or trivializing religious beliefs, symbols, or practices
Watermarks	Markings that indicate image ownership or licensure

---

📷 Image content type

i. On this project, the data will have roughly equal distribution over the categories in the first column of the table below.

Text-rich images	Examples
Documents	Books, printed articles, menus, flyers, receipts
Charts	Bar plot, pie chart, line graph, venn diagrams, tabular data
Screenshots	Mobile app, web - desktop, web-mobile
Hand-written notes	Diaries, to-do, meeting notes, study notes, letters
Text-heavy scenes	Signboards, billboards, street signs, product packaging, logos
Low-text images	**Examples**
Places	Landmarks, markets, destinations, cities, nature
People	Single/ group photos, people performing activities
Foods	Cooked meals, fruits, vegetables, packaged food items, or canned goods
Daily objects	Personal hygiene items, clothing and accessories, household items, kitchen items, technology and electronics, stationery and office supplies

Indoor scenes	Home environments, shops, buildings, offices
Outdoor scenes	Mountains, beaches, forests, cityscapes, street views
Animals	Any animal in any setting
Arts	Paintings, memes, cartoon, sculptures,

FAQs

Q1: When is it ok to split criteria?
A1: Criteria may be combined within the same challenge / request, as long as separating would not reveal a deviation in responses.

Q2: How precise must we be on prompt categories? Sometimes, they overlap 🤔
A2: Prompt categories may overlap – this is ok!

Q3: How precise must we be on rubric categories? 🧩
A3: Rubric categories may overlap–ensure the category you select for the criteria reasonably fits.

Q4: What should I do if the model's responses do not load at all? What should I do if the model's responses only partially load (i.e., one or more responses trail off mid-sentence)?
A5: If the model's responses do not load at all, resubmit the prompt. If the model's responses do load, but are only partially complete, treat the affected responses as though they were complete responses and compare them accordingly.

Criteria Categories

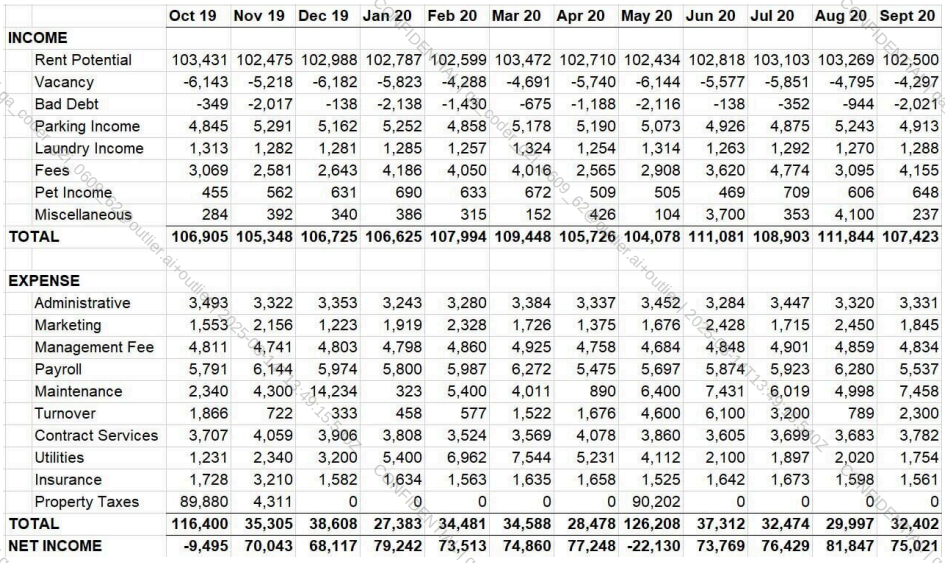
Category / Capability	Description
Instruction Following	Rubric criteria that evaluate whether the response adheres to the explicit or implied directives, constraints, or tasks provided by the user in the prompt.

Truthfulness	Rubric criteria that measure how accurately the response conveys factual, reliable information aligned with established knowledge, evidence, logical reasoning.
Writing Style / Presentation	Rubric criteria that evaluate whether the response is clearly written, reasonably concise, appropriately toned, and well-structured—ensuring readability, coherence, and visual clarity.
Content Completeness	Rubric criteria that evaluate whether the response includes all necessary and relevant information to fully satisfy the user's intent, without omitting key details that would reduce helpfulness
Visual perception (Comes from the image)	Rubric criteria that evaluate whether the response has extracted required information correctly from an image and/or used this information as necessary.
Visual reasoning (Comes from the image)	Rubric criteria that evaluate whether the response has made the right set of inferences from an image to address the prompts. The inferred information may not be directly stated in the image. For example: the period of a sine wave should be inferred from a plot based on the location of consecutive peaks.
Other	Other crucial, important, or less important criteria not captured by the above categories.

Good vs Bad Rubric Examples

Example 1

Task Type	Hyperspecific Instruction Following	
Image Category	Charts	
Prompt #1	I'm looking to make a commercial real estate investment. What kind of quantitative	

	metrics should I look out for?	
Image		

	Oct 19	Nov 19	Dec 19	Jan 20	Feb 20	Mar 20	Apr 20	May 20	Jun 20	Jul 20	Aug 20	Sept 20
INCOME												
Rent Potential	103,431	102,475	102,988	102,787	102,599	103,472	102,710	102,434	102,818	103,103	103,269	102,500
Vacancy	-6,143	-5,218	-6,182	-5,823	-4,288	-4,691	-5,740	-6,144	-5,577	-5,851	-4,795	-4,297
Bad Debt	-349	-2,017	-138	-2,138	-1,430	-675	-1,188	-2,116	-138	-352	-944	-2,021
Parking Income	4,845	5,291	5,162	5,252	4,858	5,178	5,190	5,073	4,926	4,875	5,243	4,913
Laundry Income	1,313	1,282	1,281	1,285	1,257	1,324	1,254	1,314	1,263	1,292	1,270	1,288
Fees	3,069	2,581	2,643	4,186	4,050	4,016	2,565	2,908	3,620	4,774	3,095	4,155
Pet Income	455	562	631	690	633	672	509	505	469	709	606	648
Miscellaneous	284	392	340	386	315	152	426	104	3,700	353	4,100	237
TOTAL	**106,905**	**105,348**	**106,725**	**106,625**	**107,994**	**109,448**	**105,726**	**104,078**	**111,081**	**108,903**	**111,844**	**107,423**
EXPENSE												
Administrative	3,493	3,322	3,353	3,243	3,280	3,384	3,337	3,452	3,284	3,447	3,320	3,331
Marketing	1,553	2,156	1,223	1,919	2,328	1,726	1,375	1,676	2,428	1,715	2,450	1,845
Management Fee	4,811	4,741	4,803	4,798	4,860	4,925	4,758	4,684	4,848	4,901	4,859	4,834
Payroll	5,791	6,144	5,974	5,800	5,987	6,272	5,475	5,697	5,874	5,923	6,280	5,537
Maintenance	2,340	4,300	14,234	323	5,400	4,011	890	6,400	7,431	6,019	4,998	7,458
Turnover	1,866	722	333	458	577	1,522	1,676	4,600	6,100	3,200	789	2,300
Contract Services	3,707	4,059	3,906	3,808	3,524	3,569	4,078	3,860	3,605	3,699	3,683	3,782
Utilities	1,231	2,340	3,200	5,400	6,962	7,544	5,231	4,112	2,100	1,897	2,020	1,754
Insurance	1,728	3,210	1,582	1,634	1,563	1,635	1,658	1,525	1,642	1,673	1,598	1,561
Property Taxes	89,880	4,311	0	0	0	0	0	90,202	0	0	0	0
TOTAL	**116,400**	**35,305**	**38,608**	**27,383**	**34,481**	**34,588**	**28,478**	**126,208**	**37,312**	**32,474**	**29,997**	**32,402**
NET INCOME	**-9,495**	**70,043**	**68,117**	**79,242**	**73,513**	**74,860**	**77,248**	**-22,130**	**73,769**	**76,429**	**81,847**	**75,021**

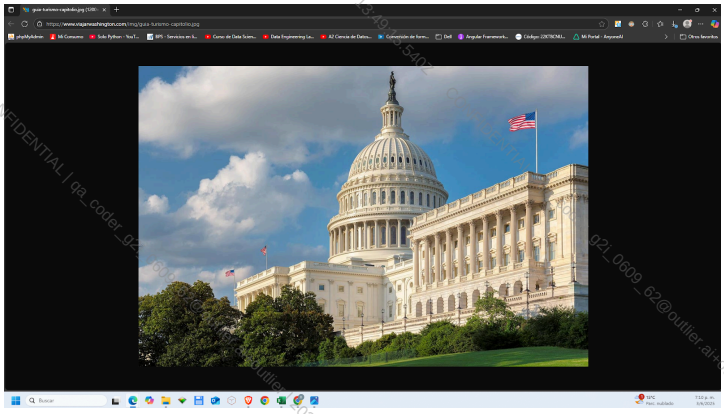
Rubric	**BAD Rubric**	**Explicit / Implicit**	**Objective / Subjective**	**Category**	**Weight**	**Explanation/Error**
	× C1: The response should state the correct industry standard assumed cap rate.	Implicit	Objective	Truthfulness	3	**[Criteria Not Self Contained]** - What is the industry standard assumed cap rate? Add “which is 6-10%”
	× C2: The response should state the annual NOI is $712,464	Explicit	× Subjective	Visual Reasoning	5	**[Criteria Objectively Wrong]** - The calculation is wrong. The sum should be 718464. **[Miscategorized Criteria]** - There is no subjectivity. This is a binary yes/no calculation
	C3: The response should state an assumed	Explicit	Objective	Instruction Following	5	

	purchased price by dividing the Annual NOI by the assumed cap rate.					
	C4: The response should state an assumed loan amount obtained by dividing the NOI by 8-10%	Explicit	Objective	Instruction Following	5	
	✕ C5: The response should state an assumed annual debt payment using a DSCR that is at least 5x.	Explicit	Objective	Truthfulness	1	**[Criteria Objectively Wrong]** - Industry average DSCR is at least 1.2, 5 doesn't make sense.
	✕ C6: The response should not show the entire calculation of NOI from the operating statement by including income/expense values for every line item.	Implicit	Objective	Writing Style / Presentation	3	**[Unnecessary criteria]** - This criteria is not critical to a golden response and is more geared to highlighting a deviation between responses.
	GOOD Rubric	**Explicit / Implicit**	**Objective / Subjective**	**Category**	**Weight**	**Error Explanation**
	✅C1: The response should state an assumed cap rate that is in a reasonable range which is 6-10%.	Implicit	Objective	Truthfulness	3	
	✅C2: The response should state the annual NOI is $718,464	Explicit	Objective	Visual Reasoning	5	
	C3: The response should state an assumed purchased price by dividing the Annual NOI by the assumed cap rate.	Explicit	Objective	Instruction Following	5	This is **explicit** because it is a quantitative metric to look out for. It's **instruction following** because it's an explicit

						directive to include quantitative metrics.
	C4: The response should state an assumed loan amount obtained by dividing the NOI by 8-10%	Explicit	Objective	Instruction Following	5	This is **explicit** because it is a quantitative metric to look out for and that is what the prompt directly asks for. It's **instruction following** because it's an explicit directive to include quantitative metrics.
	✅C5: The response should state an assumed annual debt payment using a DSCR that is at least 1.2	Implicit	Objective	Truthfulness	1	

Example 2

Task Type	Common Knowledge	
Image Category	Screenshots	
Prompt	Could you tell me what this place is and if it's possible to visit? If so, is it open to the public for tours?	

Image						
Rubric	⛔ **BAD Rubric**	**Explicit / Implicit**	**Objective / Subjective**	**Category**	**Weight**	**Error Explanation**
	C1: The response should state that the place in the image is the United States Capitol.	Explicit	Objective	Visual Perception	5 - Crucial	
	C2: The response should state it is possible to visit the United States Capitol.	Explicit	Objective	Truthfulness	5 - Crucial	
	C3: The response should explain that the Capitol is open from Monday to Saturday.	✕ Explicit	Objective	Instruction Following	3 - Important	**[Miscategorized Criteria]**- this is not explicit criteria because the prompt does not directly what days of the week the Capitol is open.
	✕ C4: The	Implicit	Objective	Truthfulness	3 -	**[Unnecessary Criteria]**- this criteria is not critical to

	response should provide relevant information, such as the necessary steps to enter the Capitol, including passing through security screening.				Important	the prompt and will cause the rubric to exceed the implicit criteria cap.
	× C5: The response should explain that the Capitol is closed on Thanksgiving, Christmas, New Year's Day, Labor Day and Inauguration Day.	Implicit	Objective	Content Completeness	3 - Important	**[Criteria Objectively Wrong] -** there is a factual inaccuracy here. The Capitol is not closed on Labor Day.
	× C6: The response should mention there are tours and that the open hours for visit tours could change for special situations and the tours typically last around 60 minutes.	Explicit	Objective	Instruction Following	5 - Crucial	**[Stacked Criteria]** - Each one of the things mentioned should be their own criteria **[Unnecessary Criteria]** - The need to mention "open hours for visit tours could change" and "tours typically last around 60 minutes" are not necessary and will cause the rubric to exceed the implicit criteria cap. Those last two points must be removed from the criteria altogether.
	C7: The response	Implicit	Objective	Truthfulness	3 - Important	

	should explain that it is always best to check the visit details on the official website which is www.visitthecapitol.gov.					
Rubric	✅ **GOOD Rubric**	**Explicit / Implicit**	**Objective / Subjective**	**Category**	**Weight**	**Error Explanation**
	C1: The response should state that the place in the image is the United States Capitol.	Explicit	Objective	Visual Perception	5 - Crucial	
	C2: The response should state it is possible to visit the United States Capitol.	Explicit	Objective	Truthfulness	5 - Crucial	
	C3: The response should explain that the Capitol is open from Monday to Saturday.	Implicit	Objective	Instruction Following	3 - Important	This is implicit criteria because the prompt does not directly ask for this information.
	C4: The response should explain that the Capitol is closed on Thanksgiving, Christmas,	Implicit	Objective	Content Completeness	3 - Important	By removing Labor Day as a holiday the Capitol is closed on this becomes a 100% factual criteria.

	New Year's Day, and Inauguration Day.					
	C5: The response should mention there are tours.	Explicit	Objective	Truthfulness	5 - Crucial	Criteria C7 from above is now “destacked”. Each individual condition has its own criteria and can measure the deviation with true precision.
	C6: The response should explain that it is always best to check the visit details on the official website which is www.visitthecapitol.gov.	Implicit	Objective	Truthfulness	3 - Important	