Coding Task Evaluation

General Overview

In Project x, you will be evaluating an AI chatbot's responses to coding-related prompts. Each evaluation will consist of two responses to the same prompt. It is crucial to identify even minor imperfections to each response. Your role will be to analyze and evaluate these responses separately, and then rate and rank the responses accordingly. You can use AI tools if you use you will be disqualify

Each evaluation task consists of five main steps:

1. You will read the prompt and the two responses to the prompt.
2. You will have to evaluate each response for how accurately it followed the instructions of the prompt, and then assign a rating to it based on this metric.
3. You will also have to evaluate each response for its overall factuality and whether all the text and code it includes is correct; you will have to test and verify the functionality and quality of all code as necessary. **For every snippet of code in the response, you will have to create and run tests to make sure it works in the given context.** Then, you will have to assign a second rating based on this metric.
4. You will write a brief explanation for each response of why you evaluated and rated each response the way that you did.
5. You will then rank one response over the other based on their overall quality (or rank them equally if they are equivalent quality) and write another explanation of your ranking choice.
6. Evaluate and Categorize Response 1 for Instruction Following - does it follow all of the instructions contained in the Prompt in terms of its goals, format, style, and type of information? Please pay close attention to detail for each item, as ALL of the prompt's specific requests must be carefully addressed in the response. Select one of the following options to categorize the quality of the response's instruction following based on the most significant issue found:
7. No Issues: The Response followed all instructions, requests, or goals in the Prompt with no missing information, steps, or data. No improvements needed;

a user would be completely satisfied with this element of the Response's quality.

8. Minor Issue(s): The Response missed or misinterpreted a small part of the instructions, requests, or goals of the Prompt. Some improvements needed; a user would be reasonably satisfied with this element of the Response's quality.
9. Major Issue(s): The Response missed or misinterpreted large parts of the Prompt's instructions, requests, or goals of the Prompt, or completely missed the point of the Prompt. Major improvements needed to prompt request adherence; a user would be mostly or completely dissatisfied with this element of the Response's quality.

1. **REQUIRED:** You **MUST add Proof of Work code** used to validate your reasoning for each snippet of code you evaluate unless the code cannot be validated within the time frame given.
    - please take a screenshot of the environment you tested the item's code snippets on, and upload it from the Evidence tab as shown in the following image.
2. If code validation is at all possible add proof of work in the given code component in the format shown in the screenshot.
3. **Make sure that every snippet of code from the response is still accounted for in your proof of work code, and add clear comments within those areas about which image you attached should be looked at for each code snippet.** If possible, explain in the proof of work code how the image shows evidence for that code snippet working or not working as intended.
4. Write a brief Explanation of Evaluation (1-3 sentences) to explain any and every error that you found for both correctness/truthfulness and instruction following. If you found none, you will still have to explain how to come to that conclusion.
5. Your explanation of your evaluation cannot directly refer / quote your Proof of Work code -instead please make it a general descriptive explanation of any errors found in the Response.
6. Repeat steps 4-8 for Response 2 before proceeding to step 10