Thank you for helping us evaluate the natural-language explanations, the final output, generated by our multi-stage explanation generation framework, MetaExplainer. We have designed the MetaExplainer, to generate explanations along five user-centered explanation types in response to user questions.

**Section 1: Metrics we are using to assess how useful MetaExplainer generated explanations are:**

Your task is to evaluate if these explanations help you achieve objectives [Hoffman et al. 2021 - <https://arxiv.org/pdf/1812.04608>] of:

1. Are you satisfied with the explanations, or do you need more details?
   1. This explanation is useful to my goals.
   2. This explanation has sufficient detail.
   3. The explanation is sufficiently complete.
   4. I understand this AI system correctly due to the explanation.
2. Did these explanations satisfy your curiosity?
   1. I want to know more about what the AI did.
   2. I want to understand what the AI will do next.
   3. I want to know why the AI did not make some other decision.
   4. I want to know what the AI would have done if something had been different.
   5. I was surprised by the AI’s actions and want to know what I missed.
3. Overall, can you trust the system better?
   1. I am confident in the MetaExplainer. I feel that it works well.
   2. The outputs of the MetaExplainer are very predictable.
   3. The MetaExplainer can perform the task better than a novice human user.
   4. I like using the system for decision making.

**Section 2: Description of evaluation material shared**

To help you assess these objectives, we include in the spreadsheet that you are provided – question, explanations in response to the question, explanation type identified to best address the question, metrics of explainer outputs included in the explanation (Tab. 1).

Tab. 1: Description of explainer metrics:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | Definition | Meaning | Modality Applied To | Source |
| Diversity |  |  | Samples |  |
| Non-representativeness |  |  | Samples |  |
| Faithfulness |  |  | Features |  |
| Monotonicity |  |  | Features |  |
| Fidelity |  |  | Rules |  |
| Average Rule Length |  |  | Rules |  |

**Your task is two-fold:**

* To fill out the multiple-choice columns that tackle Metric Set 1 against each question, explanation row. We ask that you indicate your answer as a 0: No, 1: Yes, 2: Maybe.
* On an overall scale, we would appreciate if you can score the system’s behavior based on the trust (metric set 2) and curiosity scales (metric set 3) from Sec. 1. These are in a different spreadsheet: overall system performance assessment. You will need to score these based on a Likert scale that is scaled from 1 – 5:

A white paper with black text

Description automatically generated

You will at the maximum be asked to evaluate 10 – 15 questions, note, that each participant gets a randomized set of questions from our question pool.

We will not include identifying attributes of your responses in our paper submission and instead only include a roll-up of results. The overall assessment should not take you more than 30 minutes, if it does, please feel free to reach out to Shruthi, [charis@rpi.edu](mailto:charis@rpi.edu), for more assistance.

We would appreciate if you can send us back your evaluations by this **Thursday, Aug. 8th**, preferably by noon.

**Section 3: Details of the evaluation dataset**

These results are all from a Diabetes dataset: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>, where the main objective is to predict if patients will develop type-2 diabetes in the near future or not. Hence, some explanations are providing cues post classification models being run.