**Evaluation Protocol**

1. **Objective**

The aim of this evaluation is to assess the practical effectiveness of the TRClass in classifying aerospace software requirements. Specifically, the evaluation focuses on accuracy, completeness, and helpfulness of the classification results.

1. **Participants**

* Six aerospace engineers
* At least 3 years of aerospace software development experience
* Divided equally into two groups:
  + Familiar: Participants directly involved in the project's development
  + Unfamiliar: Participants not involved in the project's development

1. **Procedure**
2. **Training Session** (30 minutes)
   * Introduction to the TRClass
   * Overview of taxonomy definitions
   * Clarification of evaluation metrics (Accuracy, Completeness, Helpfulness)
3. **Pilot Test** (30 minutes)
   * Participants conduct preliminary classifications
   * Collect feedback to improve interface clarity and refine evaluation criteria
4. **Formal Evaluation Session** (90 minutes)
   * Each participant classifies approximately 30 requirements
   * Participants use the Langchain-Chatchat-based web interface to:
     + Input or select requirements
     + View TRClass’s classification results
     + Provide structured feedback (ratings and optional comments)
5. **Data Collection**
   * Record classification queries and results
   * Collect participants’ ratings on the provided Likert scale
6. **Evaluation Criteria**

Participants will rate each classification using a 4-point Likert scale:

|  |  |
| --- | --- |
| **Rating** | **Description** |
| 1 | Disagree |
| 2 | Somewhat disagree |
| 3 | Somewhat agree |
| 4 | Agree |

Ratings cover the following dimensions:

* **Accuracy**
  + The predicted classification precisely reflects the primary intent of the requirement, avoiding ambiguous or incorrect categorizations.
* **Completeness**
  + The predicted classification includes all relevant and necessary categories without omitting critical labels.
* **Helpfulness**
  + The classification results aid participants in efficiently understanding, analyzing, or organizing the requirement, surpassing traditional manual methods.

1. **Participant Feedback**

Participants are encouraged to provide brief written comments to clarify ratings, particularly when the ratings are below 'Agree (4)', highlighting specific issues or suggestions for improvement.

1. **Post-Evaluation**

After the evaluation, results will be aggregated, analyzed, and documented to assess TRClass’s overall effectiveness and identify potential areas for enhancement.