**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.

| **Rating** | **Description** |
| --- | --- |
| 1 | Incorrect categorization; does not reflect the primary intent at all. |
| 2 | Partially incorrect; captures some aspects but misses key primary intent or includes ambiguity. |
| 3 | Mostly correct; minor inaccuracies or slight ambiguities present, but generally accurate. |
| 4 | Completely accurate; clearly and precisely matches the requirement’s primary intent without ambiguity. |

* **Completeness**
  + The predicted classification includes all relevant and necessary categories without omitting critical labels.

| **Rating** | **Description** |
| --- | --- |
| 1 | Largely incomplete; multiple critical labels missing. |
| 2 | Somewhat incomplete; at least one important label or subcategory omitted. |
| 3 | Mostly complete; includes nearly all critical labels, only minor omissions. |
| 4 | Fully complete; all relevant and necessary labels are included with no omissions. |

* **Helpfulness**
  + The classification results aid participants in efficiently understanding, analyzing, or organizing the requirement, surpassing traditional manual methods.

| **Rating** | **Description** |
| --- | --- |
| 1 | Not helpful; provides no assistance or may even hinder the analysis. |
| 2 | Somewhat helpful; slightly beneficial but not significantly more efficient than manual methods. |
| 3 | Generally helpful; noticeably improves efficiency over manual approaches, though there is some room for further improvement. |
| 4 | Highly helpful; substantially improves efficiency, clarity, and ease of analysis, clearly superior to 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.