📝 Human Evaluation Guideline for Annotators

This guideline helps human annotators evaluate AI-generated conversations involving a **teacher** and a **learner**. Each conversation should be assessed using the 8 criteria listed below. Your evaluation helps ensure high-quality instructional dialogue that is clear, engaging, and accurate.

✅ General Instructions

1. **Familiarize yourself with the conversation.** Read the full conversation carefully before scoring any criteria.

2. **Understand the roles:**

    - **Teacher:** Provides instructional content.

    - **Learner:** Asks questions and responds.

3. **Use the provided rubrics.** Each criterion has a 1–5 scale. Please reference the rubric descriptions when scoring.

4. **Be objective.** Focus on what's *actually* present in the conversation. Avoid making assumptions.

5. **Leave comments (optional but helpful).** If something stands out (positively or negatively), briefly note it.

🧭 Evaluation Criteria

1. Clarity (Teacher)

Question: Is the teacher's instruction clear, well-structured, and easy to understand?

- Look for: Clear phrasing, logical steps, no ambiguity.

- Score using the Clarity rubric from 1 (Very Poor) to 5 (Excellent).

2. Truthfulness (Teacher, with Tutorial Reference)

Question: Does the generated response stay within the scope of the reference tutorial?

- Compare the teacher's response with the **provided tutorial.**
- Identify factual accuracy and alignment with source content.
- Score from 1 (Completely Incorrect) to 5 (Perfectly Factual).

3. Engagement (Learner)

Question: Does the learner actively participate by asking meaningful questions or responding thoughtfully?

- Look for curiosity, thoughtful follow-ups, and interaction.
- Score from 1 (Very Poor) to 5 (Excellent).

4. Coherence (Conversation)

Question: Is the conversation logically structured, with smooth transitions between steps?

- Look for logical order, no abrupt topic jumps, smooth flow.
- Score from 1 (Very Poor) to 5 (Excellent).

5. Depth (Conversation)

Question: Does the conversation go beyond surface-level discussion and explore concepts in sufficient detail?

- Consider how deeply both roles explore the topic.
- Score from 1 (Very Poor) to 5 (Excellent).

6. Relevance (Conversation)

Question: Do the responses stay on-topic and align with the tutorial's instructions and context?

- Ensure responses don't stray from the tutorial goal.
- Score from 1 (Very Poor) to 5 (Excellent).

---

7. Progress (Conversation)

Question: Does the conversation effectively move forward through the tutorial steps in a logical manner?

- Check if each turn moves things forward rather than stagnating or repeating.
- Score from 1 (Very Poor) to 5 (Excellent).

---

8. Naturalness (Conversation)

Question: Does the conversation feel fluid and human-like, avoiding robotic or overly scripted responses?

- Look for a natural tone, varied phrasing, and engaging dialogue.
- Score from 1 (Very Poor) to 5 (Excellent).