Thank you for participating in this annotation task. We have provided 5 dialogues where an AI agent is responding to a user request, and we want you to evaluate the responses. This survey should take roughly 60-75 minutes. Progress will be saved automatically, so you can complete it in multiple sessions, but **make sure you are using the same browser each time.** Please rate individual dialogue responses of AI agents from 1 (worst) to 5 (best) on the following qualities: Conversation Consistency, Backend Knowledge Consistency, and Policy Compliance, the metric definitions are below. In addition, please rate the full dialogue in terms of the same metrics.

**Conversation Consistency**

How much an agent's response align with the context of conversation context.

- **Relevance**: The response directly relates to the dialogue history and the current user query.
- **Topic Consistency**: The response remains on-topic with the dialogue history and the user query.
- **Coherence**: The response logically continues the progression of the dialogue.

**Scoring Scale**

1. **Very Good (5)**: Response is completely consistent with the previous conversation context, with no inconsistencies or errors.
2. **Good (4)**: Response is mostly consistent with the context. Only minor improvements are needed.
3. **Fair (3)**: Response is somewhat consistent but contains noticeable inconsistencies or lacks depth in addressing the context.
4. **Bad (2)**: Response shows limited consistency with the conversation context and requires significant improvement.
5. **Very Bad (1)**: Response is incoherent or completely inconsistent with the conversation context.

**Backend Knowledge Consistency**

How well an agent's response aligns with information provided by backend database results.

- **Accuracy**: The response directly reflects the information in the database results.
- **Topic Consistency**: The response stays on-topic with the database results and the dialogue context.
- **Coherence**: The response logically incorporates and progresses based on the database results.

**Scoring Scale**

1. **Very Good (5)**: Response is completely consistent with the database results, with no inconsistencies or errors.
2. **Good (4)**: Response is mostly consistent with the database results. Only minor improvements are needed.
3. **Fair (3)**: Response is sufficiently consistent with the database results but contains noticeable inconsistencies or lacks depth in addressing the results.
4. **Bad (2)**: Response shows limited consistency with the database results and requires significant improvement.
5. **Very Bad (1)**: Response is incoherent or completely inconsistent with the database results.

**Policy Compliance**

How well an agent's response adheres to the expected policy protocol.

- **Number of Suggestions**: Providing suggestions only when the database results are small enough to do so.
- **Information Gathering**: Requesting required, relevant information (slots) from the user before offering suggestions or booking services.
- **Appropriate Timing**: Avoiding premature actions, such as making a booking or suggesting a service too early in the conversation.
- **Alignment with Policy**: Avoiding actions that do not align with

the suggested flow of interaction in the policy, when available.

**Expected Policy**

The chatbot response should depend on the database results and dialogue history:

- If the database results return a number larger than 10: Indicate the number of entries that match the user's query and request additional information if needed to narrow down the results.
- If the database results return values less than 10: If vital details are missing, request additional information. Otherwise, provide the relevant entries to the user

**Scoring Scale**

1. **Very Good**: Response fully follows policy protocol with no errors or omissions.
2. **Good**: Response mostly follows policy protocol, with only minor room for improvement.
3. **Fair**: Response sufficiently follows policy protocol but has clear areas where it could improve in completeness or timing.
4. **Bad**: Response does not adequately follow policy protocol, though there may be partial adherence.
5. **Very Bad**: Response fails to follow policy protocol and is incomplete or incoherent.