

# A Conversational Agent for Job Interview Practice

## A Guide For Human Evaluation

### Welcome & Thank You

Thank you for volunteering to help evaluate this conversational agent! Your feedback will help to determine whether a fine-tuned language model can better follow conversational than its pre-trained baseline. This guide explains how to carry out your evaluation task.

### TL;DR

You will have a conversation in English with a job interview practice agent. You are the candidate, and the agent is the interviewer. The conversation has 12 turns. At each turn, you will be presented two alternatives for the agent's response and a constraint. Your goal is to select the alternative that best respects the constraint.

**IMPORTANT NOTE:** *You are NOT choosing which response you prefer or which sounds better. You are choosing which response best follows a specific constraint.*

## 1. Conversational Agent Details

### 1.1. Conversation Phases

During the conversation, you will go through 5 phases. The language model automatically switches from one phase to another. Except for the first phase, it is always the agent that starts interacting in a new phase.

**Phase 1: Greetings / Small Talk** - Talking about daily life concerns, such as how was the commute to the interview place, weather, hobbies,...

**Phase 2: Background** - Start diving into the interview topic with a short discussion about your education and/or previous work experience.

**Phase 3: Technical Point** - Focus on a technical aspect related to the background you mentioned earlier. In this phase the agent will ask specific questions to evaluate your knowledge.

**Phase 4: Short Debate** - A short discussion on a controversial topic that invites to consider different perspectives. The agent will ask the question in an open way, then may challenge the point of view you expressed by suggesting counter arguments.

**Phase 5: Sum Up / Send-off** - Wrap up the interview and ask you for questions

## 1.2. Descriptors

Inside each phase, generation is controlled according to the value of one or more of the descriptors below.

- **CEFR Language Levels** - the English level of the content is assessed using the Common European Framework of Reference for Languages<sup>1</sup>. It consists in 6 levels from the easiest to the hardest:
  - A1-A2 (Basic / beginner level)
  - B1-B2 (Daily use / intermediate level)
  - C1-C2 (Advanced / advanced level)
- **Emotional Load** - whether content is emotionally charged (Loaded) or not (Non Loaded)
- **Polarity** - this feature is to be understood in the sense of sentiment analysis. It can be either Positive, Negative, or Neutral.
  - N.B. Non-neutral polarity does not necessarily imply that the content is emotionally charged. For example, the following sentence is neutral and positive: 'It is widely recognised that Italians make excellent pizzas.'

## 2. Human Evaluation Details

### 2.1. Evaluation Method

At each conversation turn, you will be given two versions of the agent's response. You will also be given the constraint that the response should follow at this stage of the conversation. Your task is to select the agent's response that you think fits best with the constraint.

Your task consists in selecting the agent response that you believe best meets the constraint.

**Reminder:** You're evaluating constraint compliance, not quality or preference! If you think there is a major quality issue in your choice's response, you should still select it and flag an inconsistency (see below).

### 2.2. Inconsistencies

Inconsistencies include all pathological cases that may occur during generation. That is, when the conversational agent does not behave as it should. After selecting a response (and only after), if you notice a particular problem, click on the 'Flag Inconsistency' button. There are three predefined types of inconsistencies (see below). If none of the three correspond to what you have identified, select 'Other' and describe the inconsistency in a few words in the dedicated text box. The predefined types of inconsistencies are:

- **Phase Inconsistency** - the agent's behavior doesn't match the expected conversation phase. Most of the time, it seems stuck in the previous phase.
- **Content Inconsistency** - the agent contradicted previous statements or forgot important context information.

---

<sup>1</sup> <https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions>

- **Nonsensical** - the generation seems broken or doesn't make any sense

## 2.3. Some Special Cases

### **Both alternatives seem equally good / are identical**

If they are not strictly identical, try to pick the one that's slightly closer to the constraint, or, if equivalent, the one that is the most relevant in the context. If both alternatives are strictly identical, pick either one.

### **One or both alternatives have the tag [EMPTY GENERATION]**

You will see this tag when the language model didn't generate anything. If only one content is empty, pick the other one anyway. If both are empty, pick either one and write another message in the following of what you have written previously. You don't need to flag an inconsistency in that case (the information of empty generation is automatically reported).

## 3. FAQ

### **Q: What if the constraint is respected but my chosen response is somehow inconsistent?**

A: First, select based on constraint compliance. Then, if you notice an inconsistency, flag it after you select the response.

### **Q: Should I flag inconsistencies in the alternative I didn't choose?**

A: No. Only flag problems in the response you selected, once you actually selected it.

### **Q: How natural should my own responses be?**

A: Respond as you naturally would in a job interview. You don't have to play your own role, you can imagine someone with a completely different background looking for a job that is not yours.

### **Q: I made a mistake in my selection. What should I do?**

A: Don't worry about it! Just continue with the next turn. There's no way to go back, and occasional errors are normal.

### **Q: How long does an evaluation session take?**

A: A complete conversation involves 12 conversation turns and takes approximately 20 minutes.

***Thank you for your participation!***