and State Detection (Obj) method outperforms the Language Only method by a small but significant margin on the user intention and step prediction tasks. Compared to Obj, the Scene Description (Sce) showed worse and more inconsistent performance across the tasks. It is likely that the visual context this method extracts is too generic or hallucinating, which on the contrary confuses the LLM predictions. We further evaluated how well these visual extraction modules perform in Section 6.3.

6.2 Instructor Decision Making

Given the context of what the user is doing and what the task environment is at each query point during task execution, the instructor needs to provide situated guidance on when to talk, and what to say in terms of intents and free-formed guidance.

We evaluate if the models can correctly predict “when to talk” at each query point, based on if there is a ground truth instructor utterance in the next x seconds, where $x = \text{median}(\text{Inst}_{speed}) \times \text{median}(\text{Inst}_{wd/uttr})$. The model could decide if it needs to talk triggered by user’s utterance, previous ground truth instructor’s comment, and visual context when no one talks. To evaluate models’ decision making performance on “what to talk” (dialog intention and free-formed guidance), we collect models’ predictions at each query condition (b), i.e. whenever the GT instructor talked. Note, the GT instructor utterances were not part of the prompt to prevent information leakage.

The overall instructor decision prediction can be found in Figure 5c. It was observed that all three methods predicted when to talk with around chance performance. Going through the examples (Figure 10, 11), we have found that ChatGPT has a stronger tendency to offer more frequent guidance when the ground truth humans don’t. All three methods have a significant above random chance aligning with the ground truth instructor and instruction intentions. Similar to user and environment understanding performances, the Language Only method demonstrated strong performance, especially in deciding the instructor’s intention, whereas the Sce method fell short across all three subtasks.

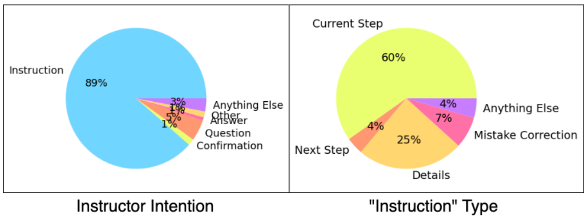

Figure 6: Dialog Intention Prediction Distribution: ChatGPT has a strong tendency to issue instructions, and especially instructions about the current step to the users. Compared to human instructors (Figure 3), the guidance is less situated, personalized, or natural.

Human language is rich and diverse and there are usually more than one acceptable way of guiding the users. Therefore, we further looked into the distributions of the model intention predictions. Comparing Figure 6 with the human intention distribution in Figure 3, the LLM tends to issue a lot more instructions to users than humans do; human instructors offer more diverse responses, including more answers and confirmation. Among all the instructions, the models tend to describe the “Current Step” whereas human instructors describe more evenly distributed instruction types. With limited user modeling and visual understanding, it is understandably harder for LLMs to offer situation-specific responses, and therefore resort to more generic instructions about the current step.

Lastly, we conducted a human evaluation on models’ generated language guidance.

Situated interactive task guidance is a personalized challenge. The guidance frequency and content vary a lot from user to user, according to their familiarity with the task itself, their chattiness, mental states, etc. We broke down the performance based on the number of dialog utterances that occurred in recordings into three categories: **short**, **mid**, and **long**. All test recordings were evenly divided accordingly. In this section, we selected 6 test recordings (2 per recipe), and asked 3 human evaluators with different genders and cultural backgrounds to rate the following for each output:

1. How helpful do you find this instructor utterance is? Rate 1/2/3, 3 = Very Helpful
2. How annoying do you think it is? Rate 1/2/3, 3 = Not Annoying

A total of 936 time points were evaluated, and each time point was rated by two annotators.

The aggregated results for the three methods are shown in Figure 7. The violin plot shows the probabilistic distribution of each rating per video category, where three horizontal lines indicate the min, max and mean ratings respectively. Overall, the average quality of the generated instructions was not considered as very helpful by the human evaluators. They are somewhat to very annoying.