# Multi-step or Direct: A Proactive Home-Assistant System Based on Commonsense Reasoning

[1,2]Konosuke Yamasaki, [3,2,1]Shohei Tanaka, [4,2]Akishige Yuguchi,
[5,2]Seiya Kawano, [6,2,1]Koichiro Yoshino

[1]Nara Institute of Science and Technology,[2]RIKEN Guardian Robot Project,[3]OMRON Sinic X
[4]Tokyo University of Science, [5]Kyoto Institute of Technology, [6]Institute of Science Tokyo
koichiro@c.titech.ac.jp

## Abstract

There is a growing expectation for the realization of proactive home-assistant robots that can assist users in their daily lives. It is essential to develop a framework that closely observes the user's surrounding context, selectively extracts relevant information, and infers the user's needs to proactively propose appropriate assistance. In this study, we first extend the Do-I-Demand dataset to define expected proactive assistance actions in domestic situations, where users make ambiguous utterances. These behaviors were defined based on common patterns of support that a majority of users would expect from a robot. We subsequently constructed a framework that infers users ' expected assistance actions from ambiguous utterances through commonsense reasoning. We explored two approaches: (1) multi-step reasoning using COMET as a commonsense reasoning engine, and (2) direct reasoning using large language models. Our experimental results suggest that both the multi-step and direct reasoning methods can successfully derive necessary assistance actions even when dealing with ambiguous user utterances.

## 1 Introduction

With recent advancements in language understanding capabilities enabled by large language models (LLMs), task-oriented domestic robots and systems that assist users based on natural language instructions are becoming a reality. Such systems interpret language (=user utterances) and map them to system actions that the system should perform, then determine an action plan based on feasible capabilities and surrounding situations (Brohan et al., 2023; Nwankwo and Rueckert, 2024; Kawaharazuka et al., 2024). The task of mapping language instruction to system actions is known as natural language understanding (NLU), and it has long been a central challenge in this field (Wang et al., 2005; Liu et al., 2021).

In NLU tasks, the user's intent is assumed to be explicit, and the primary focus has been on how to call the appropriate system action API in response to the user's utterance (Williams et al., 2016). However, real-world user utterances are often more diverse and ambiguous (Kawahara et al., 1998; Yoshino and Kawahara, 2015; Akasaki and Sassano, 2024). In such cases, the system must infer the user's potential needs or implicatures from the utterance and proactively propose supportive actions (Tanaka et al., 2024). For example, when the user says, "Thank you for the meal," it can be inferred that the user has finished eating. Ideally, the system should proactively suggest actions, such as "Shall I clear the plates on the table?," before the user explicitly makes a request.

Several challenges arise when the system tries to infer the user's potential needs from their utterances. A user's potential needs can vary greatly depending on the situation, and few proactive actions are universally acceptable to all users. The system must consider several action candidates based on commonly expected or socially appropriate actions, and select the most suitable one tailored to the users (Tanaka et al., 2021, 2023).

Moreover, when considering the appropriate proactive action in response to an ambiguous user utterance, it is often crucial to clarify the underlying assumptions behind the utterance. In this process, multi-step reasoning based on commonsense reasoning, such as COMET, is practical (Liu and Singh, 2004; Sap et al., 2019; Bosselut et al., 2019). However, multi-step reasoning without a clearly defined goal can require many inferences to reach an appropriate conclusion, and it requires a method for re-evaluating the diverse hypotheses generated during this process.

In response to these challenges, a method known as Chain-of-Thought (CoT) has recently gained attention (Wei et al., 2022). CoT involves providing LLMs with examples of rea-

*Proceedings of the 26th Annual Meeting of SIGDIAL*, pages 561–572
Aug 25–27, 2025. ©2025 SIGDIAL

soning steps, allowing them to perform inference passes (Jin et al., 2024). When using CoT to directly estimate the user's potential intent, explicit inference rules are not provided; however, the LLM can suggest plausible reasoning paths leading to the inferred conclusion. Compared to traditional multi-step reasoning approaches, CoT is expected to reduce computational cost regarding the number of inferences.

In this study, we develop a framework of proactive home-assistant robots by leveraging different reasoning methods. We extend the existing Do-I-Demand dataset (Tanaka et al., 2024) by annotating robots' support actions that are generally expected, i.e., commonsensically appropriate, by most people in response to specific user situations within a domestic setting. Using this dataset, we build an action reasoning system of proactive actions that robots should take. We compare two approaches: multi-step reasoning based on COMET and direct reasoning based on LLM, which are prompted by Chain-of-Thought (CoT) examples. Experimental results indicate that each method has its strengths and limitations, suggesting that integrating these approaches may improve accuracy in a proactive home-assistant system.

## 2 Related Works

### 2.1 Task-Oriented Dialogues and Robots

Research on domestic robots that perform assistive tasks based on natural language instructions has been actively conducted toward real-world deployment. Several approaches have been proposed, including frameworks that map natural language to executable actions (Brohan et al., 2023), multimodal robotic models that integrate vision and language (Driess et al., 2023), and end-to-end learning of language-to-action mappings using large-scale datasets (Brohan et al., 2022; O ’ Neill et al., 2024). These studies assume the user's intent is explicitly stated through direct commands, and the user's intent is linguistically clear. We seek to relax this assumption by developing a framework that infers the potential user needs from ambiguous utterances and proactively selects appropriate support actions.

### 2.2 Handling Ambiguity

Various studies have been conducted in NLU systems on robots' ability to resolve linguistic ambiguity. Some approaches integrate multimodal information, such as pointing gestures and visual context, to resolve references caused by deictic expressions or ellipses (Oyama et al., 2023; Ueda et al., 2024). Other studies have proposed methods for inferring omitted task instructions in daily life by leveraging commonsense causal relations (Takata et al., 2022). Hypothesis-driven approaches have been introduced to generate appropriate robot responses based on contextual cues, even when the user ’ s intent is not explicitly conveyed (Lanza et al., 2020). While these studies contribute to interpreting ambiguous utterances, they still assume a certain level of task-oriented intent. In contrast, this study aims to enable more autonomous and contextually "considerate" and "proactive" reasoning by interpreting not only commands or requests themselves, but also the user's situation and underlying intentions, even when the utterance itself is not directive in nature.

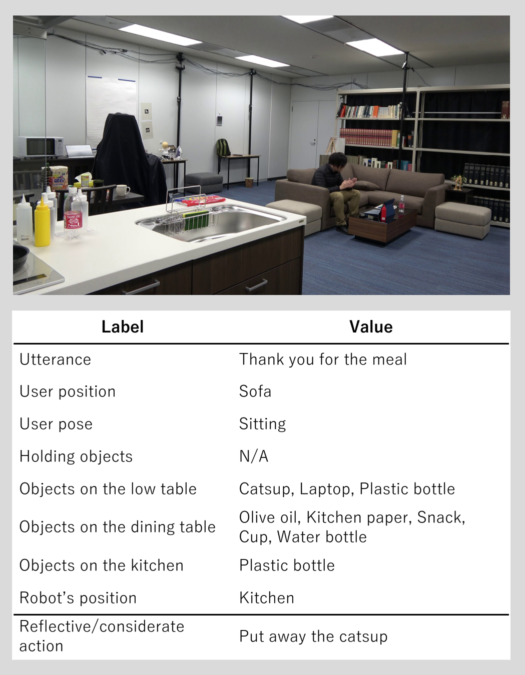
### 2.3 Reasoning and Implicature

To derive appropriate support actions from ambiguous user utterances, the system needs to perform context-sensitive reasoning and interpret implicatures embedded in the utterance (Rooy, 2001; Ruis et al., 2023). COMET is a commonsense reasoning model that captures causal and temporal relationships between everyday events as a knowledge graph and generates commonsense knowledge based on these relations (Bosselut et al., 2019). Such commonsense reasoning effectively predicts unspoken intentions or the next plausible actions. In addition, the framework of abductive reasoning, which infers the most plausible assumptions from context, helps complement the hidden intentions or goals underlying user utterances (Bhagavatula et al.). From a pragmatic perspective, research has also progressed on inferring the context-dependent meanings of utterances, beyond their surface linguistic form (Lanza et al., 2020). Building on these reasoning approaches, this study aims to infer "what the user wants the robot to do" grounded in commonsense context, and to enable robots to proactively execute appropriate actions.

## 3 Task and Dataset

### 3.1 Proactive Life Support Scenario: Do-I-Demand Dataset

To enable robots to infer and execute appropriate support actions in response to ambiguous utter-

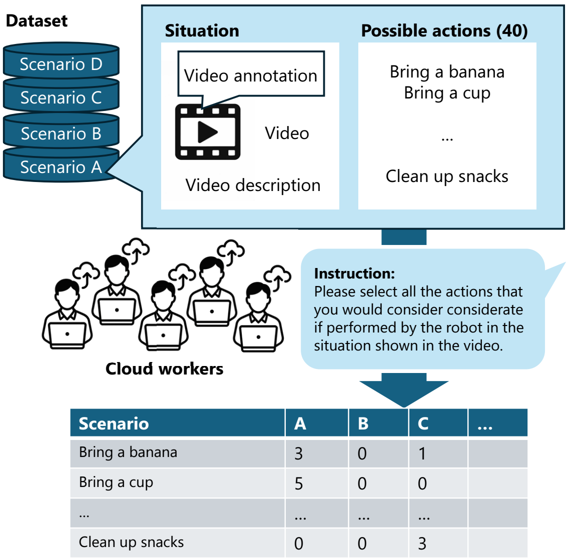


| Label | Value |
|---|---|
| Utterance | Thank you for the meal |
| User position | Sofa |
| User pose | Sitting |
| Holding objects | N/A |
| Objects on the low table | Catsup, Laptop, Plastic bottle |
| Objects on the dining table | Olive oil, Kitchen paper, Snack, Cup, Water bottle |
| Objects on the kitchen | Plastic bottle |
| Robot's position | Kitchen |
| Reflective/considerate action | Put away the catsup |

Figure 1: An example of Do-I-Demand dataset

ances, it is necessary to accurately interpret not only the utterance itself but also the underlying user intent. The Do-I-Demand dataset was constructed to realize such considerate robotic actions (Tanaka et al., 2024; Tsai et al., 2024)[1][2].

This dataset focuses on scenarios in which a domestic robot performs supportive actions, defining 40 categories of actions. For each action, pairs of ambiguous utterances and corresponding situations in which the action would be expected were collected through crowdsourcing. Workers were instructed not to give explicit commands but to describe situations where natural user utterances could be expected. Based on these descriptions, the research team recreated the situations in a real room, recorded videos, and annotated user utterances, positions, postures, and surrounding objects. Each scenario is associated with a single reflective/considerate action[3] that the robot is expected to take. Figure 1 illustrates an example. In the experiments, user utterances and visual information are used as inputs, and the prediction accuracy of deep learning models for the corresponding action is evaluated.



| Scenario | A | B | C | ... |
|---|---|---|---|---|
| Bring a banana | 3 | 0 | 1 | |
| Bring a cup | 5 | 0 | 0 | |
| ... | ... | ... | ... | |
| Clean up snacks | 0 | 0 | 3 | |

Figure 2: Our data extension procedure

While the Do-I-Demand dataset is a valuable foundation for practically evaluating reflective/considerate actions, it also presents several challenges. First, each scenario is annotated with only a single corresponding action. However, in response to ambiguous utterances, the expected action is not always uniquely determined; multiple support actions may be considered appropriate. Since users may have different expectations depending on the situation, assigning only one label fails to capture this diversity. Second, each scenario in the dataset was constructed by a crowdworker, which may introduce subjective biases. In some cases, the annotated actions do not align with the surrounding context, and there have been instances where the labeled action would be difficult or impossible for a robot to execute.

### 3.2 Multiple Action Candidates

To prevent these problems, this study conducted a reevaluation and improvement of the dataset, as shown in Figure 2. We conducted a crowdsourcing task on the original 400 scenarios, in which workers were asked to select multiple considerate actions from a set of 40 predefined action categories. Workers were provided with annotation information based on the video, including the user ’ s utterance, position, held objects, and surrounding items, and were instructed to select all applicable action categories. For each scenario, responses were collected from five workers, and the set of categories with three or more approvals (a majority) was defined as the new ground-truth label set

[1] https://github.com/riken-grp/dataloader_reflective_stretch

[2] We put our new annotations on the same repository.

[3] The existing study (Tanaka et al., 2024) defined such action as a reflective action; however, we rename this as reflective/considerate action for better understanding.

Table 1: The number of action labels per scenario

| #action | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| #scenario | 45 | 253 | 58 | 21 | 7 | 11 | 4 | 0 | 1 |

for that scenario[4].

As a result of this reevaluation, each scenario now has multiple valid action labels, allowing the dataset to better reflect the diversity of expectations regarding considerate actions. For 45 scenarios where no robot action was approved by three or more workers, the labels were revised to indicate either the absence of a valid action or an inappropriate scenario setting. The number of scenarios with at least one approved action became 355, with an average of 1.54 reflective/considerate actions per scenario. Through this reevaluation, the task has been reframed as a more realistic problem: rather than identifying a correct answer, the goal is now to flexibly select actions from multiple plausible candidate actions for a given situation.

### 3.3 Multiple Action Distribution

An analysis of the new label sets constructed through the reevaluation revealed significant variation in the number of action labels assigned to each scenario. 253 scenarios (71%) had only one action label, while the remaining 102 had two or more action labels. In some cases, up to eight actions were assigned to a single scenario, indicating that there are many situations in which the reflective/considerate action is not uniquely determined. Table 1 presents the number of actions assigned per scenario.

Many sets consisted of semantically similar actions in scenarios with multiple action labels. For example, actions such as “put away the ketchup” and “put away the cup” form a natural pair of post-meal cleanup actions. In contrast, there were cases where functionally distinct actions, such as “throw away the trash” and “bring a water bottle,” were simultaneously selected for the same situation.

Table 2: Annotated action distribution

| Action | #scenario |
|---|---|
| bring a banana | 15 |
| bring a charging cable | 12 |
| bring a cup | 18 |
| bring the ketchup | 9 |
| bring the delivery package | 6 |
| bring a plastic bottle | 23 |
| bring the remote control | 14 |
| bring the smartphone | 26 |
| bring some snacks | 23 |
| bring the tissue box | 13 |
| put away the charging cable | 8 |
| put away the cup | 37 |
| put away the ketchup | 10 |
| put away the toy car | 7 |
| put away the plastic bottle | 23 |
| put away the remote control | 5 |
| put away the smartphone | 6 |
| put away the snacks | 28 |
| put away the tissue box | 16 |
| throw away the trash | 16 |
| bring a can opener | 11 |
| bring cooking paper | 12 |
| bring a glass | 11 |
| bring a grater | 9 |
| bring kitchen paper | 11 |
| bring a lemon | 9 |
| bring olive oil | 10 |
| bring a water bottle | 24 |
| put the can opener in the cupboard | 8 |
| put the cooking paper in the cupboard | 9 |
| put the glass in the cupboard | 6 |
| put the grater in the cupboard | 8 |
| put the kitchen paper in the cupboard | 24 |
| put the plastic bottle in the fridge | 13 |
| put the plastic wrap in the cupboard | 16 |
| put the food container in the microwave | 8 |
| put the food container in the fridge | 7 |
| put the water bottle in the cupboard | 13 |

The number of scenarios in which each robot action was selected as a ground-truth label is shown in Table 2. Frequently assigned actions included “put away the cup” (37) and “put away the ketchup” (28), indicating a bias toward specific actions. In contrast, actions such as “put away the remote control” (5) and “store the grater in the cupboard” (8) were assigned less frequently. The perceived importance or frequency of these actions in daily life likely influences these biases.

We redefine the task as predicting one of the newly assigned actions for each scenario. Furthermore, we aim to improve prediction accuracy through the use of reasoning systems.

### 3.4 Analysis

We analyzed the reconstructed dataset. We found that many of the user utterances were extremely short and abstract, often lacking sufficient cues to determine appropriate actions when considered in isolation. The average number of characters per utterance was 11.74, and the average number of tokens was 5.66, with some utterances consisting of only a single word. In particular, utterances such as “I’m tired” or “It’s painful,’ are complex to link to proactive robotic actions without contextual information.

Approximately 20% of the utterances contained

[4] Our newly collected data will be publicly available when we publish this paper.

demonstratives implying context-dependent references. User utterances that included direct references to specific objects accounted for only 39%. These findings indicate that in most cases, it is difficult to identify the intended target of support based on the utterance content alone.

To address these challenges, visual contextual information, such as the user's position and posture, held objects, and surrounding items on tables or in the kitchen, can be strong cues for inferring appropriate support actions. For example, when a user says "I'm sleepy" while sitting on the sofa, the expected robot action differs from when the same utterance is made while standing in the kitchen. Moreover, when the user holds an object, it may serve as the referent of the utterance, making it an essential factor in interpreting the implied meaning. In other words, to appropriately select a reflective/considerate action in response to an ambiguous utterance, it is essential to incorporate not only the linguistic content but also complementary reasoning based on commonsense knowledge and contextual understanding of daily life. In this study, we introduce reasoning models that integrate such background knowledge to predict robotic actions in response to ambiguous user utterances.

## 4 Action Selection Based on Commonsense Reasoning

### 4.1 Inputs and Outputs

This study focuses on the task of enabling a robot to select reflective/considerate actions based on ambiguous user utterances and the surrounding context. Based on the preceding analysis, the task addressed in this work can be formalized as follows.

The input consists of the following three information:

- user utterance (uttr)
- user position (posi)
- objects that the user is holding (has)

These inputs are transformed into natural language descriptions, and the system outputs one or more relevant robotic actions from 40 predefined action classes. However, unlike conventional classification tasks, the target data in this study may correspond to multiple valid answer labels. Therefore, frameworks based on one-to-one classification or selecting a single correct answer are not applicable. Accordingly, flexible evaluation metrics that account for the set of ground-truth labels are required.

### 4.2 Evaluation Metrics

In this task, there may be multiple appropriate actions; thus, conventional metrics such as accuracy are not suitable for adequately evaluating the model's usefulness. To address this issue, we introduce the following three evaluation metrics.

**Is-in (partial matching):** This metric considers a prediction correct if at least one of the actions predicted by the model is included in the ground-truth set. It reflects scenarios where the robot has achieved its minimum objective as long as it performs at least one action that meets the user's expectations. Here, we assume that the number of actions predicted by the model is fixed[5].

$$\text{Is-in} = \frac{1}{|D|}\sum_{d\in D}\text{isin}(|P_d \cap L_d|) \qquad (1)$$

$P_d$ denotes the set of predicted actions, $L_d$ denotes the set of ground-truth actions, and $isin(x)$ is a function that returns 1 if $x \geq 1$, and 0 otherwise. $d$ is a scenario from the evaluation set $D$.

This metric assesses whether the robot can select at least one of the reflective/considerate actions defined in this dataset. However, in practice, it is also necessary to consider biases arising from differences in the importance of actions depending on the situation, as well as the frequency of the required actions themselves. This study employs the following two additional metrics in combination.

**macro-Recall:** For each scenario, we calculate the recall as the proportion of the ground-truth label set that is covered by the model ' s predictions, and then take the average across all scenarios.

$$\text{macro-Recall} = \frac{1}{|D|}\sum_{d\in D}\frac{|P_d \cap L_d|}{|L_d|} \qquad (2)$$

This metric treats scenarios with many ground-truth labels and those with only one equally, allowing us to evaluate the balance of predictions across different scenarios.

**micro-Recall:** This is the overall recall, which measures the proportion of ground-truth labels that were covered relative to the total number of

[5] We use 1, 3, and 5 in our evaluation.

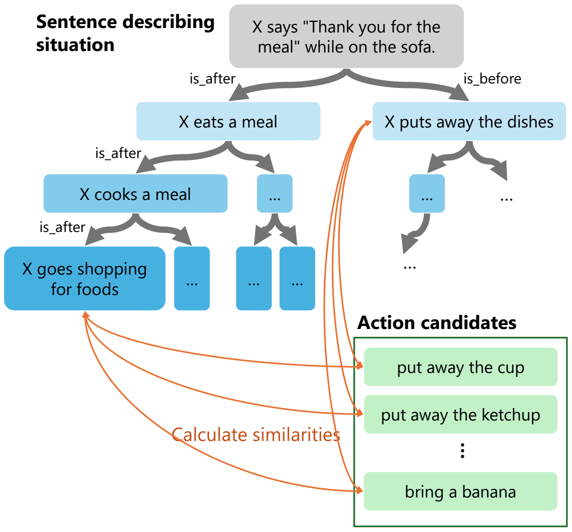


Figure 3: Using multi-step reasoning for predicting reflective/considerate actions

predicted actions.

$$\text{micro-Recall} = \frac{\sum_{d \in D} |P_d \cap L_d|}{\sum_{d \in D} |L_d|} \quad (3)$$

This metric is heavily influenced by data instances with a large number of labels, making it suitable for evaluating the overall coverage of correct actions across the entire dataset.

### 4.3 Selection Criteria

For the action prediction method, we consider approaches that can assign a ranking over the 40 candidate actions. Given the nature of the task, it is preferable to select multiple candidate actions rather than just one. To determine the predicted action set $P_d$ from this ranked list, we introduce the following two settings:

- Top-$k$: Use the top $k$ actions from the ranked list as $P_d$ (e.g., Top-1, Top-3)
- Plus-$k$: Select the top-ranked actions up to the number of ground-truth labels plus $k$ for each data instance (e.g., in the Plus-1 setting, if a scenario has two ground-truth actions, the top three actions from the ranked list are selected as $P_d$.)

### 4.4 Multi-step Reasoning using COMET

To determine reflective/considerate actions in response to ambiguous utterances, it is often necessary to perform multi-step reasoning of the user's intent and their possible future actions. In this section, we propose a method that uses commonsense knowledge based on an event knowledge graph to perform multi-step reasoning. The system selects reflective/considerate actions based on the semantic similarity between the final reasoning result and the candidate actions, as shown in Figure 3.

COMET is an inference engine trained on a large-scale event knowledge graph called ATOMIC2020. In this study, we use COMET-ja, a Japanese-language variant of COMET[6]. Given an input event (e.g., "X says 'Thank you for the meal' while on the sofa"), it outputs related events that are likely to occur before or after the input (e.g., -(is_after)-> "X eats a meal", -(is_before)-> "X puts away the dishes"). The specific procedure is outlined below:

**Multi-step reasoning:** We input a sentence into COMET that integrates the user ' s utterance (uttr) and contextual information, such as position (posi) and held object (has). COMET then generates events that represent possible subsequent actions. These generated events are used recursively to perform inference up to a maximum depth of $T$. At each inference step, the previous event is used as input, and COMET outputs events connected by a specified relation.

**Calculating similarities:** We compute the semantic similarity between all generated events and the 40 candidate actions. We use the Japanese version of SimCSE (Tsukagoshi et al., 2023) along with cosine distance. For each action candidate, the highest similarity score among the generated events is used as its score, and the actions are ranked based on these scores. The action with the highest score is selected as the reasoning result for the ambiguous utterance.

**Inference depth-based penalty:** Increasing the maximum inference depth $T$ allows for generating a broader range of related events; however, it also increases the risk of logical leaps. To address this issue, we introduce a depth-based penalty $\alpha_t$ in this study. The following calculation is used:

$$S_m(l_{d,i}, e_j) = Sim(l_{d,i}, e_j) \prod_{t=1}^{T} \alpha_t \quad (4)$$

Here, $l_{d,i}$ denotes a candidate action, $e_j$ is an inferred event, and $Sim(\cdot)$ represents the similarity function. The value of $\alpha_t$ is determined using a leave-one-out method based on the other scenarios. The proposed multi-hop reasoning approach

[6] We used comet-v2-gpt2-small-japanese from https://github.com/nlp-waseda/comet-atomic-ja

using COMET has the advantage of providing an explicit reasoning process, thereby enhancing the explainability of the robot's action.

### 4.5 Direct Reasoning by LLM

Another approach for reasoning reflective/considerate actions for ambiguous utterances is direct end-to-end reasoning using large language models (LLMs). In this section, we describe a method that directly generates appropriate actions for a given user situation by leveraging the extensive prior knowledge and contextual understanding capabilities of LLMs, without relying on conventional multi-step reasoning.

In this approach, contextual information is provided as input, such as the user's utterance (uttr), position (posi), and holding objects (has). A prompt is constructed in which the model is asked to enumerate all appropriate actions from among the 40 predefined action candidates. We design a 1-shot prompt for OpenAI ' s LLMs as follows:

- Present all 40 executable robot actions in advance
- Explicitly describe the user's situation (uttr, posi, and has)
- Instruct the model to select all appropriate action categories and output them using their action numbers

The LLM can leverage its internal language knowledge and contextual understanding to selectively extract appropriate actions with this prompt design. Since the LLM enumerates plausible actions based on the given situation, we rank the candidate actions according to the order in which they are output by the model. The used prompt is shown in the appendix.

We use GPT-3.5 and GPT-4 as the OpenAI models in this study. Since the knowledge used to train COMET is based on GPT-3, using GPT-3.5 allows for a fair comparison of the effectiveness of multi-step reasoning compared to direct reasoning. However, we also include GPT-4 to assess the state-of-the-art performance on this task.

The primary advantage of this direct reasoning approach is that it does not require the explicit construction of a knowledge graph or the design of multi-step reasoning procedures. LLMs can generate reasonable outputs by capturing the overall context of a situation. Furthermore, because much of the implicit commonsense knowledge is embedded within LLMs, they can perform flexible reasoning without relying on specific external knowledge sources. On the other hand, this approach has limitations, such as difficulty in ensuring output consistency and explainability, as well as in strictly controlling task-specific constraints.

## 5 Results

### 5.1 Experimental Settings

In this section, we conduct a comparative evaluation of the proposed methods (Multi-step and Direct) and the two baseline methods (Similarity and Random). We use the evaluation metrics defined in Section 4.2—Is-in, macro recall, and micro recall—and evaluate each method under the Top-k (k = 1, 3, 5) and Plus-k (k = 0, 1, 3, 5) settings.

As baselines for comparison with the proposed method, we define the following two approaches: (1) Similarity baseline, based on semantic similarity; (2) Random ranking, which assigns action rankings at random. (1) computes the semantic similarity between the user's utterance (including contextual information) and each action candidate, and generates a ranking based on these similarity scores. (2) uses a randomly shuffled list of the 40 action candidates as the ranking. This baseline serves as a lower bound for performance and offers a reference point for evaluating the task's difficulty.

### 5.2 Experimental Results

Table 3 shows the Is-in results for each method and evaluation condition. For variations within the same method, we report only the setting that achieved the best score due to space limitations.

Among the baselines, the similarity-based method using only the utterance achieved the highest score. For the Multi-step method, performance improved as the inference depth increased when using only uttr. However, when using uttr + posi + has, performance decreased as inference depth increased. These results suggest that effective utilization of multi-step inference requires careful design of both the input information for reasoning and the depth-based penalty applied during reasoning.

The Direct approach achieved the higher overall scores. A comparison between Direct (GPT-3.5) and Multi-step shows that, with appropriately tuned parameters, the multi-step approach can outperform the Direct approach. The results for Top-

Table 3: Scores of **Is-in**. We are highlighting the first and the second-best scores.

| Method | Top-1 | Top-3 | Top-5 | Plus-0 | Plus-1 | Plus-3 | Plus-5 |
|---|---|---|---|---|---|---|---|
| Multi-step ($T$=1, uttr+posi+has) | 0.341 | 0.583 | 0.676 | 0.420 | 0.552 | 0.656 | 0.707 |
| Multi-step ($T$=3, uttr) | **0.473** | **0.724** | **0.786** | **0.549** | **0.701** | **0.780** | **0.837** |
| Direct (GPT-3.5) | 0.454 | 0.648 | 0.673 | 0.518 | 0.620 | 0.676 | 0.696 |
| Direct (GPT-4) | **0.727** | **0.808** | **0.839** | **0.769** | **0.820** | **0.839** | **0.845** |
| Similarity (uttr) | 0.454 | 0.707 | 0.766 | 0.532 | 0.693 | 0.766 | 0.828 |
| Random | 0.054 | 0.115 | 0.175 | 0.090 | 0.121 | 0.180 | 0.234 |

Table 4: Scores of **macro-Recall**

| Method | Top-1 | Top-3 | Top-5 | Plus-0 | Plus-1 | Plus-3 | Plus-5 |
|---|---|---|---|---|---|---|---|
| Multi-step ($T$=1, uttr+posi+has) | 0.304 | 0.506 | 0.586 | 0.340 | 0.468 | 0.571 | 0.618 |
| Multi-step ($T$=3, uttr) | 0.408 | 0.622 | 0.684 | 0.449 | 0.598 | 0.685 | 0.742 |
| Similarity (uttr) | 0.395 | 0.615 | 0.671 | 0.435 | 0.594 | 0.673 | 0.731 |
| Random | 0.034 | 0.079 | 0.115 | 0.044 | 0.068 | 0.111 | 0.160 |

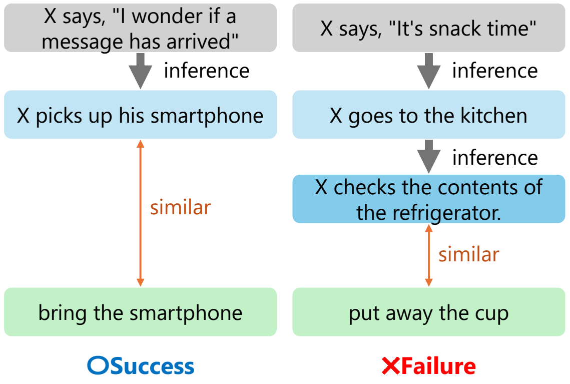

X says, "I wonder if a message has arrived"
inference
X picks up his smartphone
similar
bring the smartphone
OSuccess
X says, "It's snack time"
inference
X goes to the kitchen
inference
X checks the contents of the refrigerator.
similar
put away the cup
✕Failure

Figure 4: Success/failure cases of multi-step reasoning

5 and Plus-5 indicate that the multi-step reasoning approach can prioritize more appropriate action candidates in the ranking. However, the performance of the multi-step approach is highly sensitive to parameter settings, and its increased complexity does not always justify the gains in performance. Moreover, Direct (GPT-4) improved significantly even over the Multi-step, confirming that recent advancements in LLM performance have greatly contributed to this task.

Tables 4 and 5 show the results for **macro-Recall** and **micro-Recall**, respectively. The trends observed in these results were consistent with those for **Is-in**. This indicates that, overall, the dataset constructed in this study exhibits minimal bias in terms of the importance of actions depending on the situation and the frequency of the required actions.

Figure 4 shows success/failure examples of the multi-step reasoning method. In the success case, the model correctly associates the event of receiving a message with the appropriate action involving a smartphone. In the failure case, although the model appropriately associates the phrase “snack time” with “refrigerator,” the distance-based evaluation metric fails to select the correct action. Another possible explanation is that the distance metric used in this study (SimCSE) may not be entirely suitable for this task. This point warrants further investigation in future work.

## 6 Conclusion

In this study, we developed a framework for estimating appropriate support actions for domestic assistive robots, based on ambiguous user utterances and contextual information, to realize reflective/considerate actions that users expect from such systems. We first re-evaluated and extended the existing Do-I-Demand dataset. By relabeling the original single reflective/considerate action with a set of actions that are widely agreed upon as appropriate by multiple annotators, we enabled more flexible evaluation. Our dataset reflects the real-world assumption that user expectations are not always uniquely determined by utterances. We also proposed two distinct reasoning approaches to predict such reflective/considerate actions. The multi-step reasoning method used a commonsense inference system (COMET) to generate related events from the utterance and infer actions. The direct inference approach using LLMs directly predicted actions by jointly processing the utterance and contextual information. Experimental results demonstrate that both proposed methods outperformed the similarity-based baseline. The multi-step reasoning method outperformed GPT-3.5, which uses a knowledge base comparable to COMET.

When we use the multi-step reasoning approach, it is crucial to carefully select the knowl-

Table 5: Scores of **micro-Recall**

| Method | Top-1 | Top-3 | Top-5 | Plus-0 | Plus-1 | Plus-3 | Plus-5 |
|---|---|---|---|---|---|---|---|
| Multi-step ($T$=1, uttr+posi+has) | 0.221 | 0.399 | 0.475 | 0.297 | 0.397 | 0.490 | 0.543 |
| Multi-step ($T$=3, uttr) | 0.307 | 0.497 | 0.574 | 0.397 | 0.512 | 0.605 | 0.669 |
| Similarity (uttr) | 0.294 | 0.486 | 0.559 | 0.380 | 0.505 | 0.592 | 0.658 |
| Random | 0.035 | 0.077 | 0.121 | 0.064 | 0.086 | 0.132 | 0.181 |

edge fed into the reasoning process, determine which relations to use, and design effective pruning strategies for generated event candidates. The selection method may also be improved through enhanced similarity measures. Since the information in the real world that needs to be attended to varies depending on the target action, future work should explore methods that better use knowledge specifically related to the action candidates.

## Limitation

This paper determines reflective/considerate actions through majority voting based on situational context. However, it does not include individual user studies, and therefore it is not possible to evaluate whether the selected actions were truly appropriate for each individual user. Taking actions that users perceive as reflective or considerate requires considering a broader context, not only the surrounding situation, but also user preferences and the long-term relationship between the robot and the user.

## Acknowledgments

This work was supported by JST Moonshot Project Grant Number JPMJMS2236.

## References

Satoshi Akasaki and Manabu Sassano. 2024. Detecting ambiguous utterances in an intelligent assistant. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 386–394.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. 2022. RT-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. 2023. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on robot learning*, pages 287–318. PMLR.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. PaLM-E: An embodied multimodal language model. In *International Conference on Machine Learning*, pages 8469–8488. PMLR.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1830–1842.

Tatsuya Kawahara, Chin-Hui Lee, and Biing-Hwang Juang. 1998. Flexible speech understanding based on combined key-phrase detection and verification. *IEEE transactions on speech and audio processing*, 6(6):558–568.

Kento Kawaharazuka, Tatsuya Matsushima, Andrew Gambardella, Jiaxian Guo, Chris Paxton, and Andy Zeng. 2024. Real-world robot applications of foundation models: A review. *Advanced Robotics*, 38(18):1232–1254.

Davide Lanza, Roberto Menicatti, and Antonio Sgorbissa. 2020. Abductive recognition of context-dependent utterances in human-robot interaction. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10975–10981. IEEE.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. Benchmarking natural language understanding services for building conversational agents. In *Increasing naturalness and flexibility in spoken dialogue interaction: 10th international workshop on spoken dialogue systems*, pages 165–183. Springer.

Linus Nwankwo and Elmar Rueckert. 2024. The conversation is the command: Interacting with real-world autonomous robots through natural language. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 808–812.

Akira Oyama, Shoichi Hasegawa, Hikaru Nakagawa, Akira Taniguchi, Yoshinobu Hagiwara, and Tadahiro Taniguchi. 2023. Exophora resolution of linguistic instructions with a demonstrative based on real-world multimodal information. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 2617–2623. IEEE.

Abby O ' Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE.

Robert Rooy. 2001. Conversational implicatures. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, 36:20827–20905.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Kota Takata, Takuya Kiyokawa, Ixchel G Ramirez-Alpizar, Natsuki Yamanobe, Weiwei Wan, and Kensuke Harada. 2022. Efficient task/motion planning for a dual-arm robot from language instructions and cooking images. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12058–12065. IEEE.

Shohei Tanaka, Konosuke Yamasaki, Akishige Yuguchi, Seiya Kawano, Satoshi Nakamura, and Koichiro Yoshino. 2024. Do as i demand, not as i say: A dataset for developing a reflective life-support robot. *IEEE Access*.

Shohei Tanaka, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Arta: Collection and classification of ambiguous requests and thoughtful actions. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 77–88.

Shohei Tanaka, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Reflective action selection based on positive-unlabeled learning and causality detection model. *Computer Speech & Language*, 78:101463.

Shang-Chi Tsai, Seiya Kawano, Angel-Garcia Contreras, Koichiro Yoshino, and Yun-Nung Chen. 2024. ASMR: Augmenting life scenario using large generative models for robotic action reflection. In *In Proceedings of International Workshop on Spoken Dialogue Systems Technology (IWSDS) 2024*.

Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. 2023. Japanese simcse technical report. *arXiv preprint arXiv:2310.19349*.

Nobuhiro Ueda, Hideko Habe, Akishige Yuguchi, Seiya Kawano, Yasutomo Kawanishi, Sadao Kurohashi, and Koichiro Yoshino. 2024. J-CRe3: A japanese conversation dataset for real-world reference resolution. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9489–9502.

Ye-Yi Wang, Li Deng, and Alex Acero. 2005. Spoken language understanding. *IEEE Signal Processing Magazine*, 22(5):16–31.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Koichiro Yoshino and Tatsuya Kawahara. 2015. News navigation system based on proactive dialogue strategy. *Natural language dialog systems and intelligent assistants*, pages 15–25.

## A Used prompt

Tables 6 and 7 show the used prompt for direct reasoning by GPT, and their translation.

Table 6: Used prompt with an example

| role | content |
|---|---|
| system | あなたはリビングに居るユーザの状況から、生活支援のためのロボットが実行する気が利く行動を考えるアシスタントです。<br>気が利く行動は、明示的に指示されてはいないがユーザに役立つ行動です。<br>与えられるユーザの状況に対して、最後に示す 40 個の行動カテゴリの中からロボットが実行する行動として適切な気が利く行動を全て選択してください。<br>これらの行動カテゴリはどのような状況でも全て実行することができます。<br>回答は、行動カテゴリとその行動カテゴリに割り振られた番号だけを出力し、各行動カテゴリごとに改行してください。<br>適当な行動カテゴリが 1 つも無い場合には「該当なし」とだけ出力してください。<br>回答が終了したら [回答終了] と出力してください。<br>以下は 40 個の行動カテゴリの一覧です。<br>----------<br>[1] バナナを持ってくる<br>[2] 充電ケーブルを持ってくる<br>[3] コップを持ってくる<br>[4] ケチャップを持ってくる<br>[5] 宅配便を持ってくる<br>[6] ペットボトルを持ってくる<br>[7] リモコンを持ってくる<br>[8] スマホを持ってくる<br>[9] お菓子を持ってくる<br>[10] ティッシュ箱を持ってくる<br>[11] 充電ケーブルを片付ける<br>[12] コップを片付ける<br>[13] ケチャップを片付ける<br>[14] ミニカーを片付ける<br>[15] ペットボトルを片付ける<br>[16] リモコンを片付ける<br>[17] スマホを片付ける<br>[18] お菓子を片付ける<br>[19] ティッシュ箱を片付ける<br>[20] ゴミをゴミ箱に入れる<br>[21] 缶切りを持ってくる<br>[22] クッキングシートを持ってくる<br>[23] グラスを持ってくる<br>[24] おろし器を持ってくる<br>[25] キッチンペーパーを持ってくる<br>[26] レモンを持ってくる<br>[27] オリーブオイルを持ってくる<br>[28] じゃがいもを持ってくる<br>[29] サランラップを持ってくる<br>[30] 水筒を持ってくる<br>[31] 缶切りを棚にしまう<br>[32] クッキングシートを棚にしまう<br>[33] グラスを棚にしまう<br>[34] おろし器を棚にしまう<br>[35] キッチンペーパーを棚にしまう<br>[36] ペットボトルを冷蔵庫にしまう<br>[37] サランラップを棚にしまう<br>[38] タッパーをレンジに入れる<br>[39] タッパーを冷蔵庫にしまう<br>[40] 水筒を棚にしまう<br>---------- |
| user | 【ユーザの状況】<br>ユーザの発話: どこに片付けたかな？<br>ユーザの位置: キッチン<br>ユーザが手にしている物: 大根<br>ユーザの近くにある物: 無し |
| assistant | [24] おろし器を持ってくる<br>[回答終了] |
| user | 【ユーザの状況】<br>ユーザの発話: これだけで朝までもつかな？<br>ユーザの位置: ソファ<br>ユーザが手にしている物: 無し<br>ユーザの近くにある物: パン, ペットボトル |

Table 7: Used prompt with an example (Translation)

| role | content |
|---|---|
| system | You are an assistant that helps a robot determine considerate actions to support daily life, based on the user's situation in the living room.<br>A considerate action is one that is not explicitly instructed but is helpful to the user.<br>Given the user's situation, please select all appropriate considerate actions the robot should take from the list of 40 action categories shown at the end.<br>All of these categories can be performed in any situation.<br>Your response should output only the action category and its assigned number, one per line.<br>If none of the categories are appropriate, output only "None applicable".<br>After completing the response, output [End of response].<br>Below is the list of 40 action categories:<br>----------<br>[1] Bring a banana<br>[2] Bring a charging cable<br>[3] Bring a cup<br>[4] Bring ketchup<br>[5] Bring a delivery package<br>[6] Bring a plastic bottle<br>[7] Bring a remote control<br>[8] Bring a smartphone<br>[9] Bring snacks<br>[10] Bring a tissue box<br>[11] Put away the charging cable<br>[12] Put away the cup<br>[13] Put away the ketchup<br>[14] Put away the toy car<br>[15] Put away the plastic bottle<br>[16] Put away the remote control<br>[17] Put away the smartphone<br>[18] Put away the snacks<br>[19] Put away the tissue box<br>[20] Throw trash into the bin<br>[21] Bring a can opener<br>[22] Bring cooking sheet<br>[23] Bring a glass<br>[24] Bring a grater<br>[25] Bring kitchen paper<br>[26] Bring a lemon<br>[27] Bring olive oil<br>[28] Bring a potato<br>[29] Bring plastic wrap<br>[30] Bring a water bottle<br>[31] Put the can opener in the cabinet<br>[32] Put the cooking sheet in the cabinet<br>[33] Put the glass in the cabinet<br>[34] Put the grater in the cabinet<br>[35] Put the kitchen paper in the cabinet<br>[36] Put the plastic bottle in the refrigerator<br>[37] Put the plastic wrap in the cabinet<br>[38] Put the container in the microwave<br>[39] Put the container in the refrigerator<br>[40] Put the water bottle in the cabinet<br>---------- |
| user | [User Situation]<br>User utterance: Where did I put it away?<br>User location: Kitchen<br>Item in user's hand: Daikon radish<br>Items near the user: None |
| assistant | [24] Bring a grater<br>[End of response] |
| user | [User Situation]<br>User utterance: I wonder if this will be enough to last until morning.<br>User location: Sofa<br>Item in user's hand: None<br>Items near the user: Bread, plastic bottle |