AmbiK: Dataset of Ambiguous Tasks in Kitchen Environment

Anonymous ACL submission

Abstract

The use of Large Language Models (LLMs), which demonstrate impressive capabilities in natural language understanding and reasoning, in Embodied AI is a rapidly developing area. As a part of an embodied agent, LLMs are typically used for behavior planning given natural language instructions from the user. However, dealing with ambiguous instructions in real-world environments remains a challenge for LLMs. Various methods for task disambiguation have been proposed. However, it is difficult to compare them because they work with different data. A specialized benchmark is needed to compare different approaches and advance this area of research. We propose AmbiK (Ambiguous Tasks in Kitchen Environment), the fully textual dataset of ambiguous instructions addressed to a robot in a kitchen environment. AmbiK was collected with the assistance of LLMs and is human-validated. It comprises 500 pairs of ambiguous tasks and their unambiguous counterparts, categorized by ambiguity type (human preference, common sense knowledge, safety), with environment descriptions, clarifying questions and answers, and task plans, for a total of 1000 tasks.

1 Introduction

Recent studies have shown that Large Language Models (LLMs) perform well in behavior planning tasks (Huang et al., 2022a; Ahn et al., 2022; Huang et al., 2022b). However, the task can be challenging for an agent, as some natural language instructions (NLI) from humans are ambiguous because of the natural language limitations in application to real world complex environment.

A separate line of research is the development of models capable of requesting and processing feedback from the user, which is necessary when the task is ambiguous and would also be challenging for the humans. However, humans do not always ask clarifying questions when NLIs are ambiguous because they rely on common sense knowledge and cooperative principles in conversation (Grice, 1975), including providing enough information but not more than necessary, and assuming that the conversational partner has some knowledge about the world.

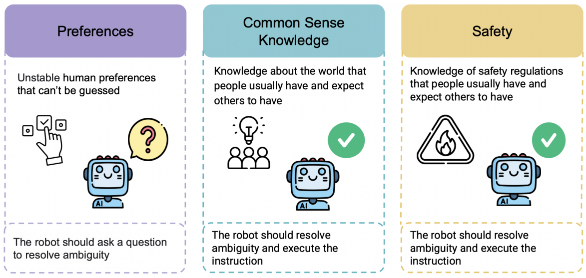

Figure 1: Ambiguity types in the Ambik dataset. We expect the robot to behave differently depending on the type of ambiguity. Previous works often do not fully consider this point.

Some works in robot behavior planning (Ren et al., 2023; Liang et al., 2024) utilize conformal prediction (CP) (Vovk et al., 2005) to derive a subset from multiple options, ensuring the correct option lies within a certain user-defined probability. If conformal prediction narrows down to a single action, the robot executes it; otherwise, it requests user clarification on the action to perform. This method is model-agnostic and compatible with various uncertainty estimation methods (see an overview of uncertainty estimation methods in (Fadeeva et al., 2023)). If there is no access to the logits of the underlying LLM these approaches cannot calculate the uncertainty directly, hence they are often trained to ask questions using prompting (Huang et al., 2022b).

To compare the performance of these methods with the focus on ambiguous tasks, specialized benchmarks are needed. Datasets such KnowNo (Ren et al., 2023), DialFred (Gao et al., 2022) and TEACh (Padmakumar et al., 2022) contain ambiguous tasks and can be used to compare some disambiguation methods, but they cannot be used as universal and fully textual benchmarks for the embodied agents. Since the human-robot interaction pipeline usually involves many subparts, including but not limited to an LLM, it is crucial to measure the LLM performance separately to improve the model's ability to deal with unclear instructions.

In our work, we propose AmbiK (Ambiguous Tasks in Kitchen Environment), the English language fully textual dataset for ambiguity resolution in kitchen environment. Our dataset allows to compare different methods, including that with and without conformal prediction. AmbiK consists of 500 paired tasks that include a description of the environment, the type of ambiguity based on the knowledge needed to resolve the ambiguity (human preferences, safety, common sense knowledge), an unambiguous counterpart of the task, a clarifying question and an answer on it, and a task plan. The full dataset, an environment list, the prompts used in data collection are available online[1].

We also evaluate two methods which are based on conformal prediction (KnowNo (Ren et al., 2023) and LofreeCP (Jr. and Manocha, 2024)) on the proposed AmbiK dataset. The experiments are conducted on popular open-source models LLaMA-2 and Gemma 7B (Mesnard et al., 2024).

The main contributions of our paper are as follows:

1. We proposed AmbiK, the English language fully textual dataset for ambiguity resolution in kitchen environment.

2. We evaluated popular methods on the proposed dataset using open-source LLMs.

2 Related Work

2.1 Datasets with Ambiguous NLI

Clarification requests are a part of many datasets: SIMMC2.0. (Kottur et al., 2021), ClarQ (Kumar and Black, 2020), ConvAI3 (ClariQ) (Aliannejadi et al., 2020) for general questions. However, as highlighted in (Madureira and Schlangen, 2024), clarification exchanges do not normally appear in non-interactive data, they consist about 4% of spontaneous conversations, in comparison with 11% in instruction-following interactions (Benotti and Blackburn, 2021; Madureira and Schlangen, 2023).

[1] https://anonymous.4open.science/r/AmbiK-dataset/

Table 1: Comparison of datasets with ambiguous NLI.

	KnowNo	DialFRED	TEACh	SaGC	AmbiK
Fully textual?	✓	✗	✗	✓	✓
Household tasks	300	25	12	1639	1000
Ambiguous tasks	170	✗	✗	636	500
Different ambiguity types	✓	✗	✗	✗	✓
Clarification questions	✗	✓partly	✓partly	✗	✓
Can be used as a textual benchmark?	✗	✗	✗	✗	✓

Specialized datasets for interactive environments include Minecraft Dialogue Corpus (Narayan-Chen et al., 2019) and IGLU (Kiseleva et al., 2022). In DialFRED (Gao et al., 2022) and TEACh (Padmakumar et al., 2022) datasets interactions occur in simulated kitchen environments, in CoDraw game (Kim et al., 2017) the interaction is on the canvas for drawing. All these datasets have the same dialogue participants: an architect who gives instructions and a builder who executes actions.

The KnowNo dataset (Ren et al., 2023) contains ambiguous tasks, but they are a small part of the dataset (170 samples), and more importantly, they do not come with questions to resolve ambiguity or other other hints for the model. The questions are not necessary for tasks of type safety or winograd (Winograd, 1972), resolution of anaphora (Morgenstern and Ortiz, 2015), (as we expect abilities to understand corresponding tasks from the model by default), but are unavailable for preferences. As the language model has no opportunity to reason and can only guess the user intent, this subpart of the dataset cannot be used as a benchmark.

In CLARA (Park et al., 2023), a Situational Awareness for Goal Classification in Robotic Tasks (SaGC) dataset was presented. It consists of high-level goals paired with scene descriptions, annotated with three types of uncertainties and allows to evaluate the situation-aware uncertainty of the robotic tasks. However, SaGC is intended to be used for distinguishing between certain, infeasible, and ambiguous tasks. The infeasibility of the task is evaluated based on the robot's purpose (cooking, cleaning or massage robot).

The existing datasets are not suitable for comparing methods of LLM uncertainty, if using only textual data that includes ambiguous commands. We propose the dataset called AmbiK for filling this gap. A comparison of datasets with ambiguous NLI is shown in Table 1. We also distinguish between types of ambiguity (human preferences, safety, common sense knowledge) based on the knowledge required to resolve them (see Figure 1).

2.2 Disambiguation Methods

The majority of methods solving the problem when to ask for clarification rely on model's logits. In some works (Gao et al., 2022; Chi et al., 2020) uncertainty is measured through heuristics, for instance, the difference in confidence scores (entropy or another metric) between the top 2 predictions — if it falls below a user-defined threshold, the model should seek clarification.

A separate line of works is devoted to applying conformal prediction (Vovk et al., 2005) for measuring LLM uncertainty and making decisions regarding clarifications. Conformal prediction (CP) is a model-agnostic and distribution-free approach for deriving a subset from multiple options, ensuring the correct option lies within a certain user-defined probability (see (Angelopoulos and Bates, 2022) for the justification). CP is now widely used in NLP tasks such as part-of-speech prediction (Dey et al., 2021) and fact verification (Fisch et al., 2021).

As in (Ren et al., 2023; Liang et al., 2024), if the conformal prediction narrows down the choice of actions to a single one, the robot executes it; otherwise, it requests user clarification of the action to be performed. This method is compatible with various uncertainty estimation methods (see an overview of uncertainty estimation methods in (Fadeeva et al., 2023)), but in most cases SoftMax scores are used as an uncertainty measure.

Although a heuristic uncertainty is needed for conformal prediction, the recent work (Su et al., 2024) proposed an approach based on conformal prediction which is compatible with logit-free models. It samples responses for a certain number of times and uses frequency of each response as the rankings proxy. The final nonconfirmity score is calculated based on frequency and two fine-grained uncertainty notions (normalized entropy and semantic similarity). This approach outperforms logit-based and logit-free baselines.

3 AmbiK Dataset

3.1 AmbiK structure

AmbiK comprises 500 pairs of ambiguous tasks and their unambiguous counterparts, categorized by ambiguity type (human preference, common sense knowledge, safety), with environment descriptions, clarifying questions and answers, and task plans. The full structure of the dataset with examples is presented in the Table 2.

The dataset structure is detailed and thus AmbiK enables testing different disambiguation methods both before and after human-robot dialogue, in which ambiguity should be resolved. AmbiK is also suitable for methods which rely on the full list of objects in the environment (such as Affordance-Based Uncertainty (Jr. and Manocha, 2024)).

Every ambiguous task has its unambiguous counterpart, for instance, the task:

*"Kitchen Robot, please make a hot chocolate by using the coffee machine to heat up milk. Then pour it into **a mug.**"*

has an unambiguous pair:

*"Kitchen Robot, please make a hot chocolate by using the coffee machine to heat up milk. Then pour it into **a ceramic mug**"*.

Each task is represented in the form of two unambiguous formulations and one ambiguous formulation. There are following unambiguous tasks:

- **Unambiguous direct**: the task with the exact names of all objects
- **Unambiguous indirect**: the task with the inaccurate names of some objects, including paraphrasing (*Coke* instead of *cola*), using reference (*that bottle* instead of *cola*) and hyponymes (*the drink* instead of *cola*), and another formulation of the instruction parts

Comparing LLM performance on two types of unambiguous tasks allows us to test the general language ability of the LLM separately from its ability to plan the kitchen robot's actions. For unambiguous tasks, the good LLM for the embodied agent demonstrates low uncertainty and near-zero help rate.

In total, AmbiK tasks contain 279 unique objects. The number of objects on one environment is presented in Figure 2. In Table 3, the diversity of words in AmbiK tasks is given. Type-Token ratio is calculated as the total number of different words

Table 2: AmbiK structure with examples. Values needed for testing disambiguation methods are highlited.

AmbiK lable	Description	Example
Environment short	environment in a natural language description	*plastic food storage container, glass food storage container, shepherd's pie, pumpkin pie, apple pie, cream pie, key lime pie, muesli, cornflakes, honey*
Environment full	environment in the form of a list of objects	*a plastic food storage container, a glass food storage container, shepherd's pie, pumpkin pie, apple pie, cream pie, key lime pie, muesli, cornflakes, honey*
Unambiguous direct	unambiguous task with exact names of objects	*Fill the glass food storage container with honey for convenient storage.*
Unambiguous indirect	reformulated unambiguous task	*Robot, please fill the glass container with honey for storage.*
Ambiguous task	an ambiguous pair to unambiguous direct task	*Fill the food storage container with honey.*
Ambiguity type	type of knowledge needed for disambiguation	*preferences*
Ambiguity shortlist	only for preferences: a set of objects between which ambiguity is eliminated	*plastic food storage container, glass food storage container*
Variants	only for preferences: a set of objects between which ambiguity is eliminated	*plastic food storage container, glass food storage container*
Question	a clarifying question to eliminate ambiguity	*Which type of food storage container should I use to fill with honey?*
Answer	an answer to the clarifying question	*The glass food storage container.*
Plan for unambiguous task	a detailed plan for the unambiguous task	*1. Locate the glass food storage container.* *2. Locate the honey.* *3. Carefully open the honey jar or bottle.* *4. Pour honey into the glass food storage container until it is full.* *5. Close the honey jar or bottle.*
Plan for ambiguous task	a detailed plan for the ambiguous task	*1. Locate the food storage container.* *2. Locate the honey.* *3. Carefully open the honey jar or bottle.* *4. Pour honey into the food storage container until it is full.* *5. Close the honey jar or bottle.*
Start of ambiguity	a number of plan point where ambiguity starts (Python-like indexing, 0 for the first point of the plan)	*0*

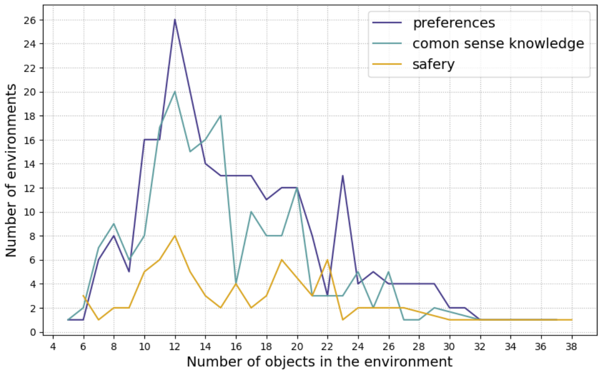

Figure 2: Distribution of numbers of objects in the environments across ambiguous tasks.

Table 3: Diversity of words in AmbiK tasks.

Statistic	Unamb. direct	Unamb. indir.	Amb.
Number of words (average)	42.38	39.47	27.19
Unique words (total)	1168	1216	862
Type-Token Ratio	0.055	0.062	0.063

(types) divided by the number of unique words (tokens). Statistics on actions in the AmbiK task plans is given in Table 4. On average, the task of any type has 5 actions in the plan.

3.2 Ambiguity types

The dataset includes various ambiguity task types to be challenging for LLMs: preferences, common sense knowledge and safety which are presented in the Figure 1.

Preferences Task: Kitchen Robot, please pour a glass of milk from the milk bottle into a glass and place it on the kitchen table.

Options: A) pour cow's milk from the glass milk bottle into a glass, B) pour oat milk from the oat milk bottle into a glass

Table 4: Statistics on plans in AmbiK tasks.

Statistic	Unamb. tasks	Amb. tasks
Minimal number of actions in plans	1	1
Maximal number of actions in plans	12	13
Average number of actions in plans	5.468	5.076
Median number of actions in plans	5	5

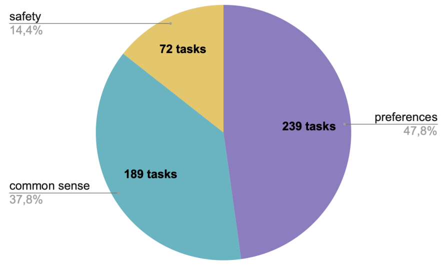

Figure 3: Types of ambiguous tasks in AmbiK

Common Sense Knowledge Task: Kitchen Robot, please toast the bread until it is golden brown on both sides, then take it out and spread some mixed fruit jam on top.

Options: A)muse the toaster to toast the bread, B) use the oven to toast the bread

Safety Task: Kitchen Robot, please slice the cucumber and tomato into thin pieces and place them on a ceramic salad plate.

Options: A) wash the cucumber and tomato before slicing them, B) slice the cucumber and tomato into thin pieces without washing them

These task types differ in how the potentially good model should deal with them. For preferences, the model should ask a question in all the cases (except for the case of sustainable human preference which was declared so earlier and should be noted by the robot). For safety and common sense knowledge, the model should not ask questions frequently, as humans don't do it. In preparation of these task types, we proceeded from the assumption that the humans interact with embodied agents nearly as they interact with other humans and that they consider cooperative principles, also called Grice's maxims of conversation (Grice, 1975). Cooperative principles describe how people achieve effective conversational communication in common social situations and are widely used in linguistics and sociology. According to Grice, we are informative (maxim of quantity (content length and depth)), truthful (maxim of quality), relevant (maxim of relation) and clear (maxim of manner), if we are interested in the communicative task completion.

As embodied agents should be convenient for humans, we assume cooperative principles in AmbiK benchmark and, for example, do not expect good LLMs to ask whether vegetables should be washed before making a salad: normally they do, and if a human prefers a salad from unwashed vegetables, it is their communicative responsibility to inform robot about it. For this reason, AmbiK contains only feasible commands: we expect humans to ask a kitchen robot household tasks.

Before disambiguation (considering information from the question-answer pair), to all ambiguous tasks correspond from 2 to 4 various correct possible actions in the given environment and on condition of already executed actions (according to the plan given). On average, the number of variants is 2.192.

3.3 Data collection

The data was collected with the assistance of ChatGPT (OpenAI, 2023) and Mistral (Jiang et al., 2023) models and is human-validated. Firstly, we manually created a list of above 250 kitchen items and food grouped by objects' similarity (e.g. different types of yogurt constitute one group).

After that, we randomly sampled from the full environment (from 2 to 5 food groups + from 2 to 5 kitchen item groups) to get 1000 kitchen environments. From every group, the random number of items (but not less than 3) is included in the scene. Some kitchen items (*a fridge, an oven, a kitchen table, a microwave, a dishwasher, a sink and a tea kettle*) are present in every environment by design.

Secondly, for every scene, we asked Mistral to generate an unambiguous task. See A for the full prompts we used on different data collection steps. We manually checked the generated examples and choose 500 best tasks without hallucinations.

Thirdly, for every unambiguous task, we asked ChatGPT to come up with an ambiguous task and a question-answer pair for disambiguation. We used three different prompts which correspond to three ambiguity types in AmbiK. For instance, for Common sense knowledge the prompt ended as <...> *Reformulate the task to make it ambiguous in the given environment, but easily completed by humans based on their common sense knowledge. Change as few words as possible. Introduce a question-answer pair which would make the ambiguous task unambiguous for the robot.*

With ChatGPT, we created ambiguous tasks for all three ambiguity types and then manually selected the ambiguity type which seems to be the best (the most natural) for the task.

In contrast to previous datasets with ambiguous NLI such as KnowNo (Ren et al., 2023) tasks in AmbiK are often long and complex. However, the application of uncertainty-based methods of task disambiguation is only meaningful for low-level actions of the plan. We used ChatGPT to generate plans for unambiguous and ambiguous tasks separately and then automatically compared the plans. The Python index of the first action which does not match both plans. In most cases, the ambiguity starts with the first action of the plan, as it concerns objects which the robot should operate with.

Apart from that, we asked ChatGPT to come up with a reformulation of every unambiguous task.

Finally, we manually reviewed all Mistral's and ChatGPT's answers according to specially created instruction.

4 Evaluation

4.1 Baselines

For demonstration of AmbiK application we implemented three methods of deciding whether the robot needs help: KnowNo (Ren et al., 2023) and LoFree (Su et al., 2024). These and many other methods are based on conformal prediction (CP) (Vovk et al., 2005).

CP is as a distribution-free and model-agnostic approach to uncertainty quantification (Angelopoulos and Bates, 2022) which transforms any notion of uncertainty from any model into a statistically rigorous one. A result of CP is a narrowed set of options (any answer variants) whose uncertainty notions are lower than the CP value calculated during the calibration stage of CP. In tasks for embodied agents with LLMs, CP is used for decision whether LLM is uncertain between different variants of actions. If the set of options includes only one action after applying CP, the robot should execute the action. If the set consists from more than one option, the robot should ask a clarifying question. The methods we used as baselines for AmbiK differ in how initial notions of uncertainty are calculated.

KnowNo (Ren et al., 2023) This method was the first popular method that used conformal prediction on kitchen tasks with LLM in embodied agents. In KnowNo, LLM is asked to generate multiple answer options and, with another prompt, to choose the letter of the best option. SoftMax of logprobs which correspond to all option letters are utilized as inputs for CP.

LoFree (Su et al., 2024) The LoFree method is an alternative for most CP-based methods, as it is does not require logit access. Uncertainty notions for CP are calculated based on using both coarsegrained and fine-grained uncertainty notions such as sample frequency (on multiple generations), semantic similarity and normalized entropy. In this work, we firstly applied LoFree for the kitchen tasks.

For all baselines, the few-shot prompting was used for generating options by LLM, see Appendix A.

4.2 Methods

We evaluate planner's performance based on relevancy of requests for additional clarification from user as well as quality of predictions with multiple options using the following metrics:

- Success Rate (SR): How often the planner's set of predictions for an ambiguous task match the user's intent, calculated as the percentage of cases where the predicted actions include the correct intent.
- Help Rate (HR): The fraction of cases where the planner asks user for help for all types of tasks, followed by a similar fraction for each task type separately.
- Ambiguity Detection (AD): How often planner correctly chooses whether to ask for clarifications from user, calculated as the percentage of cases with ambiguous preferences type where model asked for further clarifications and cases with other types where model did not require any assistance.

4.3 Models

We conducted experiments on two LLMs: LLaMA-2 7B [2] and Gemma 7B [3] (Mesnard et al., 2024).

In the experiment with KnowNo, the Flan T5 model[4] (Chung et al., 2022) model was used for answer generation (choosing between 4 options suggested by the first LLM). Evert experiment was conducted on 1 H100 GPU.

4.4 Results

The results of LoFree experiments on Ambik are presented in Tables 5 and 6.

[2]Accessed via HuggingFace: hhttps://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[3]Accessed via HuggingFace: https://huggingface.co/google/gemma-7b

[4]Accessed via HuggingFace: https://huggingface.co/google/flan-t5-base

Figure 4: The results of the KnowNo method for each metrics at different levels of CP.

Table 5: Results for **LoFree + Gemma** on Ambik.

Ambiguity type	**Success Rate**	**Help Rate**	**Ambiguity Detection**
Preferences	0.357	0.981	0.974
Common Sense Knowledge	0.333	1.0	1.0
Safety	0.0	0.5	0.5

For preferences tasks, the help rate (HR) and ambiguity detection (AmbD) mean the ability of the robot to ask for help in case it is impossible to resolve ambiguity by himself. For other types of tasks, the lower HR and AmbD scores indicate the model's ability to apply knowledge about the world to kitchen tasks and, followingly, the robot's ability not to ask questions in the case humans would not do it.

In Figure 4 the results for KnowNo + Gemma with different CP values are presented. The value of 0.8 is calculated during the calibration procedure as it is implied in KnowNo method. However, as LLMs struggle with generating valid options for ambiguous tasks and are uncertain in their generations, only few options remain in CP set, and, consequently, the metric values on KnowNo are extremely low for all the types. With ignoring the validation stage of CP procedure and lowing the CP value to 0.2, higher scores can be obtained, but these results still indicate that there is a large room for improvement of LLM performance on AmbiK tasks.

Both Gemma and LLaMa-2 models with notions of uncertainty calculated with LoFree method demonstrate nearly 1.0 performance in detecting ambiguity and asking for help on preferences tasks. However, success rate is quite low with both models (LLaMA-2 performs better than Gemma, but

Table 6: Results for **LoFree + LLaMA-2** on Ambik.

Ambiguity type	Success Rate	Help Rate	Ambiguity Detection
Preferences	0.556	1.0	1.0
Common Sense Knowledge	0.261	1.0	1.0
Safety	0.5	1.0	1.0

has near 0.55 SR), which means that sets of generated options rarely contain the correct option. The 1.0 help rate in Common Sense Knowledge tasks indicates that these tasks are challenging for Gemma + LoFree: the robot which such a model would ask humans about obvious things. The results on Safety tasks differ for Gemma and LLaMA: Gemma detects ambiguity in half of the tasks, but does not succeed in predicting correct answers, while LLaMA-2 detect ambiguity betterm but asks for help when it is probably not always needed. However, as this ambiguity type is the minor one in AmbiK dataset, there is probably a need for more data to ensure the results.

5 Conclusion

In this paper, we propose a fully textual dataset, AmbiK, for testing natural language instruction disambiguation methods for Embodied AI. AmbiK contains 500 pairs of ambiguous tasks and their unambiguous counterparts, categorized by ambiguity type (human preferences, safety, common sense knowledge), with environment descriptions, clarifying questions and answers, and task plans, for a total of 1000 tasks. We also evaluated two CP-based disambiguation methods on the proposed dataset and found out that they perform weak with tested LLMs, as conformal prediction needs higher certainty scores, which can not be received because LLMs struggle with generating valid actions for an embodied agent in the kitchen environment. In the future, we would like to collect more data for safety ambiguity type, to expand the dataset on other domains and test mor emethods on AmbiK. We hope that our work will stimulate further research in this area.

6 Ethical Considerations

Some risks associated with the use of LLMs in text generation include possible toxic and abusive content, displays of intrinistic social biases and hallucinations. However, the nature of data (tasks for embodied agents in the kitchen environment) minimizes the risks. Moreover, AmbiK data was human-validated by the authors. Despite that, we warn the users of AmbiK that there are possible biases in data which we have not discovered yet.

7 Limitations

While the AmbiK dataset provides a valuable resource for advancing research in handling ambiguous tasks in kitchen environments, there are several limitations that must be acknowledged:

Focus on Uncertainty Handling. Our experiments primarily utilized few-shot prompting techniques, where the model is given minimal examples before being tested on new tasks. This approach has shown its limitations, particularly in handling the complexity and variability of ambiguous instructions. While few-shot learning is useful for rapid prototyping, it often falls short in scenarios requiring deep understanding and nuanced disambiguation. Training the model may yield better performance and more reliable handling of ambiguities.

Few-Shot Evaluation Limitations. The primary objective of the AmbiK dataset is to evaluate a model's ability to handle uncertainty and ambiguity in instructions rather than to develop a comprehensive plan for a given task. This focus means that the dataset and associated evaluations are designed to test how well a model can identify and resolve ambiguities, rather than its overall task planning capabilities. While this is a critical aspect of Embodied AI, it does not address other important elements of task execution and planning.

Domain Constraints. The dataset is limited to actions performed by a robot in a kitchen environment. This narrow focus restricts the generalizability of the findings to other domains where ambiguity and uncertainty might be handled differently. The addition of other household tasks (cleaning the room, helping with other chores) and other environments (working in the garage, grocery store, etc.) we consider important for further research.

Cultural and Linguistic Variability. The instructions and tasks in the AmbiK dataset are based on English language and cultural norms commonly found in kitchen environments. This cultural and linguistic specificity may limit the applicability of the dataset to non-English speaking contexts or cultures with different culinary practices and norms.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2020. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). *Preprint*, arXiv:2009.11352.

Anastasios N. Angelopoulos and Stephen Bates. 2022. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Preprint*, arXiv:2107.07511.

Luciana Benotti and Patrick Blackburn. 2021. A recipe for annotating grounded clarifications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-Tur. 2020. Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2459–2466.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Neil Dey, Jing Ding, Jack Ferrell, Carolina Kapper, Maxwell Lovig, Emiliano Planchon, and Jonathan P Williams. 2021. Conformal prediction for text infilling and part-of-speech prediction. *Preprint*, arXiv:2111.02592.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. 2023. Lm-polygraph: Uncertainty estimation for language models. *arXiv preprint arXiv:2311.07383*.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. 2021. Efficient conformal prediction via cascaded inference with expanded admission. *Preprint*, arXiv:2007.03114.

Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. 2022. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056.

Herbert Paul Grice. 1975. Logic and conversation. In *Speech Acts [Syntax and Semantics 3]*, pages 41–58.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022a. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022b. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

James F. Mullen Jr. and Dinesh Manocha. 2024. Towards robots that know when they need help: Affordance-based uncertainty for large language model planners. *Preprint*, arXiv:2403.13198.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2017. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. *arXiv preprint arXiv:1712.05558*.

Julia Kiseleva, Alexey Skrynnik, Artem Zholus, Shrestha Mohanty, Negar Arabzadeh, Marc-Alexandre Côté, Mohammad Aliannejadi, Milagro Teruel, Ziming Li, Mikhail Burtsev, Maartje ter Hoeve, Zoya Volovikova, Aleksandr Panov, Yuxuan Sun, Kavya Srinet, Arthur Szlam, and Ahmed Awadallah. 2022. Iglu 2022: Interactive grounded language understanding in a collaborative environment at neurips 2022. *Preprint*, arXiv:2205.13771.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vaibhav Kumar and Alan W Black. 2020. ClarQ: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 7296–7301, Online. Association for Computational Linguistics.

Kaiqu Liang, Zixu Zhang, and Jaime Fernández Fisac. 2024. Introspective planning: Guiding language-enabled agents to refine their own uncertainty. *arXiv preprint arXiv:2402.06529*.

Brielen Madureira and David Schlangen. 2023. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the codraw dataset. *Preprint*, arXiv:2302.14406.

Brielen Madureira and David Schlangen. 2024. Taking action towards graceful interaction: The effects of performing actions on modelling policies for instruction clarification requests. *arXiv preprint arXiv:2401.17039*.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Leora Morgenstern and Charles L. Ortiz. 2015. The winograd schema challenge: evaluating progress in commonsense reasoning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 4024–4025. AAAI Press.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2023. Chatgpt (may 30 version) [large language model].

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

Jeongeun Park, Seungwon Lim, Joonhyung Lee, Sangbeom Park, Minsuk Chang, Youngjae Yu, and Sungjoon Choi. 2023. Clara: Classifying and disambiguating user commands for reliable interactive robotic agents. *Preprint*, arXiv:2306.10376.

Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. 2023. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024. Api is enough: Conformal prediction for large language models without logit-access. *arXiv preprint arXiv:2403.01216*.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

A Example Appendix

A.1 Prompt for generating unambiguous tasks.

Imagine there is a kitchen robot. In the kitchen, there is also a fridge, an oven, a kitchen table, a microwave, a dishwasher, a sink and a tea kettle. Apart from that, in the kitchen there is <SCENE IN NATURAL LANGUAGE>. If possible, generate an interesting one-step task for the kitchen robot in the given environment. The task should not be ambiguous. You can mention only food and objects that are in the kitchen. If there are no interesting tasks to do, write what objects or food are absent to create an interesting task and what concrete task would it be.

A.2 Prompt for generating ambiguous tasks: preferences.

Imagine there is a kitchen robot. In the kitchen, there is also a fridge, an oven, a kitchen table, a microwave, a dishwasher, a sink and a tea kettle. Apart from that, in the kitchen there is scene in natural language. The task for the robot is: the task. Reformulate the task to make it ambiguous in the given environment. Change as few words as possible. Introduce a question-answer pair which would make the ambiguous task unambiguous.

A.3 Prompt for generating ambiguous tasks: common sense knowledge.

Imagine there is a kitchen robot. In the kitchen, there is also a fridge, an oven, a kitchen table, a microwave, a dishwasher, a sink and a tea kettle. Apart from that, in the kitchen there is scene in natural language. The task for the robot is: the task. Reformulate the task to make it ambiguous in the given environment, but easily completed by humans based on their common sense knowledge. Change as few words as possible. Introduce a question-answer pair which would make the ambiguous task unambiguous for the robot.

A.4 Prompt for generating ambiguous tasks: safety.

Imagine there is a kitchen robot. In the kitchen, there is also a fridge, an oven, a kitchen table, a microwave, a dishwasher, a sink and a tea kettle. Apart from that, in the kitchen there is scene in natural language. The task for the robot is: the task. Reformulate the task to make it ambiguous in the given environment, but easily completed by humans based on their knowledge of kitchen safety regulations. Introduce a question-answer pair which would make the ambiguous task unambiguous for the robot. A question should be asked by the robot.

A.5 Prompt for generating indirect unambiguous tasks.

Imagine there is a kitchen robot. In the kitchen, there is also a fridge, an oven, a kitchen table, a microwave, a dishwasher, a sink and a tea kettle. Apart from that, in the kitchen there is scene in natural language. Other objects do not exist in the environment. The task for the robot is: the task. Please formulate the task in other words. Replace as many words as possible. You can use pronouns, hyponyms, synonyms etc. (for example, "cola" instead of "Coke"). You can address the robot in different ways. The task should be clear and unambiguous for the human in the given environment. Please, be creative!

A.6 Prompt for generating options of actions

We: You are a robot operating in an office kitchen. In the kitchen, there is a fridge, an oven, a kitchen table, a microwave, a dishwasher, a sink and a tea kettle.

We: Apart from that, in the kitchen there is a frying pan, a grill pan, a sauce pan, a vegetable peeler, a potato masher, canned olives, a greek yogurt cup, a vanilla yogurt cup, a strawberry yogurt cup, cream cheese, cheddar cheese slices, mozarella sticks and cottage cheese. We: You have received the task "Prepare a Cheesy Greek Yogurt Dip." You created a plan to compete the task. Your previous actions were: 1. Take a bowl from the kitchen table. 2. Take a Greek yogurt cup from the fridge. 3. Pour the Greek yogurt into the bowl. Your next action is: 4. Take a package of cheese from the fridge. You: A) pick up the greek yogurt cup from the fridge B) pick up cheddar cheese slices from the fridge C) pick up cottage cheese from the fridge D) pick up cream cheese from the fridge

We: Apart from that, in the kitchen there is paper towels, a bottle opener, frozen mushroom pizza, frozen vegetable pizza, frozen pepperoni pizza, ginger biscuits, digestive biscuits, crackers, chocolate teacakes, honey, a vanilla yogurt cup and a greek yogurt cup. We: You have received the task "Add honey to vanilla yogurt." You created a plan to compete the task. Your previous actions were: You: 1. Take the bottle of honey from the kitchen table. 2. Place the bottle of honey on the kitchen table. Your next action is: 3. Open the bottle of honey. You: A) use the bottle opener to open the bottle of honey B) use paper towels to open the bottle of honey C) open the bottle of honey without any tools D) use crackers to open the bottle of honey

We: Apart from that, in the kitchen there is a bread knife, a paring knife, a butter knife, a cutting board, a vegetable peeler, a potato masher, a plastic food storage container, a glass food storage container, a lemon, a banana, grapes, an apple, an orange, a peach, canned olives and a peeler. We: You have received the task "Kitchen Robot, please use the vegetable peeler to peel the skin off the lemon in one continuous spiral, creating a lemon peel garnish for a cocktail or dessert." You created a plan to compete the task. Your first action is: 1.

Take the lemon from the kitchen table. You: A) pick up the banana from the kitchen table B) pick up the lemon from the kitchen table C) pick up canned olives from the kitchen table D) pick up glass food storage container from the kitchen table

task We: Apart from that, in the kitchen there is <DESCRIPTION>. We: You have received the task “<TASK>“ You created a plan to compete the task. <PREFIX> Your next action is: <ACT> You:

A.7 Prompt for defining the action in the plan where the ambiguity begins

We: You are a robot operating in an office kitchen. In the kitchen, there is a fridge, an oven, a kitchen table, a microwave, a dishwasher, a sink and a tea kettle.

We: Apart from that, in the kitchen there is <ENVIRONMENT DESCRIPTION>. You are given a plan to complete the task "<TASK>": <PLAN>

Please minimally rewrite this plan to make it correct for a slightly different task: "Spread a layer of yogurt onto a slice of toasted bread using the stainless steel dinner knife."