AmbiK: Dataset of Ambiguous Tasks in Kitchen Environment

Anastasiia Ivanova
MIPT, HSE University
Moscow, Russia

Alexey K. Kovalev
AIRI, MIPT
Moscow, Russia

Aleksandr I. Panov
AIRI, MIPT
Moscow, Russia

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. To address this issue and further advance this area of research, a specialized benchmark is needed. We propose AmbiK, the fully textual dataset of ambiguous commands addressed to a robot in a kitchen environment. AmbiK was collected with the assistance of LLMs and is human-validated. It comprises 250 pairs of ambiguous tasks and their unambiguous counterparts, categorized by ambiguity type, with additional information, for a total of 500 tasks.

1. Introduction

Recent studies have shown that Large Language Models (LLMs) perform well in behavior planning tasks [1, 7, 8]. 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. Studies indicate that a model's ability to ask clarifying questions based on LLM uncertainty enhances task performance [6, 19]. Some works in robot behavior planning [14, 21] utilize conformal prediction [22] 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 [5]). Models without open logs cannot directly calculate uncertainty, hence they are often trained to ask questions using prompting [8].

Table 1. Comparison of datasets with ambiguous NLI.

	KnowNo	DialFRED	TEACh	SaGC	AmbiK
Fully textual?	✓	✗	✗	✓	✓
Household tasks	300	25	12	1639	500
Ambiguous tasks	170	✗	✗	636	250
Different ambiguity types	✓	✗	✗	✗	✓
Clarification questions	✗	✓partly	✓partly	✗	✓
Can be used as a textual benchmark?	✗	✗	✗	✗	✓

To compare the performance of these methods with the focus on ambiguous tasks, specialized benchmarks are needed. Datasets such KnowNo [21], DialFRED [6] and TEACh [19] contain ambiguous tasks and can be used to compare some diasmbiguation 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 ambiguous instructions.

We propose AmbiK (Ambiguous Tasks in Kitchen Environment), the 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 250 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].

2. Datasets with Ambiguous NLI

Clarification requests are a part of many datasets: SIMMC2.0. [12], ClarQ [13], ConvAI3 (ClariQ) [3] for general questions. However, as highlighted in [16], 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 [4, 15]. Specialized datasets for interactive environments include Minecraft Dialogue Corpus [18] and IGLU [11]. In DialFRED [6] and TEACh [19] datasets interactions occur in simulated kitchen environments, in CoDraw game [10] 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 [21] 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, resolution of anaphora [17], (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 [20], 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 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.

3. AmbiK dataset

Data collection. The data was collected with the assistance of ChatGPT [2] and Mistral [9] models and is human-validated. Firstly, we manually created a list of above 130 kitchen items and food and sampled it to get 1000 kitchen environments. Some kitchen items (such as a microwave, a fridge, etc.) are present in every environment by design. Secondly, we asked Mistral to come up with an interesting unambiguous task for the kitchen robot in the given environment or write which items are absent for a possible interesting task. Thirdly, we manually checked the generated examples and choose 250 best tasks without hallucinations. After that, for every task, we generated an ambiguous counterpart and a question-answer pair for task disambiguation using ChatGPT. We created ambiguous tasks for every ambiguity type, manually selected the best ambiguity type and created a list of items which are in the scope of ambiguity. ChatGPT was also prompted to come up with creative reformulations of tasks on the condition that the task remains unambiguous.

Dataset structure. The dataset has the following structure (simplified examples are in italics):

1. environment in a natural language description *(a microwave, a plastic bowl, and a metal bowl)*
2. environment in the form of a list of objects *(microwave, plastic bowl, metal bowl)*
3. task *(Place the bowl in the microwave, please)*
4. ambiguity type: preferences, common sense knowledge, safety
5. a set of objects between which ambiguity is eliminated *(plastic bowl, metal bowl)*
6. a clarifying question to eliminate ambiguity *(Which bowl should I take?)* and an answer *(A plastic bowl.)*
7. plan *(1. Pick up plastic bowl. 2. Go to the microwave. 3. Place plastic bowl in the microwave.)*

Task types. The dataset consist of various task types to be challenging for LLMs. **Unambiguous tasks**: 1. direct tasks (with the exact name of objects), 2. indirect tasks (with the inaccurate name of objects, e.g. paraphrasing *(Coke instead of cola)*, reference *(that bottle)*, hyponymes *(the drink)*, etc.). Two types of unambiguous tasks let test the general language ability of LLMs. **Ambiguous tasks:** 1. preferences (human preferences), 2. safety (knowledge of safety regulations), 3. common sense knowledge (common sense knowledge about the world, e.g. knowledge of what objects are commonly used for – the task *"wash it and put it on the table"* hardly applies to a microwave or chips).

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 ... Then pour it into* ***a ceramic mug****"*.

4. Conclusion

In this paper, we propose a fully textual dataset, **AmbiK**, for testing disambiguating natural language instructions methods for Embodied AI. AmbiK comprises 250 pairs of ambiguous tasks and their unambiguous counterparts, categorized by ambiguity type, with environment descriptions, clarifying questions and answers, and task plans, for a total of 500 tasks. As further research, we consider extending the proposed dataset both by adding more examples and by using other environments in addition to the kitchen.

[1] https://github.com/cog-model/AmbiK-dataset/

Acknowledgements

The results were obtained with the support of Sberbank (№ 70-2021-00138) which is part of the program and plan of research center in the field of AI provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement on the provision of subsidies (identifier of the agreement 000000D730324P540002) and the agreement with the Moscow Institute of Physics and Technology dated November 1, 2021 No. 70-2021-00138.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1

[2] Open AI. Chatgpt (may 3 version) [large language model]. 2023. 2

[3] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq), 2020. 2

[4] Luciana Benotti and Patrick Blackburn. 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, 2021. 2

[5] Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. Lm-polygraph: Uncertainty estimation for language models. *arXiv preprint arXiv:2311.07383*, 2023. 1

[6] Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056, 2022. 1, 2

[7] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022. 1

[8] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022. 1

[9] 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. Mistral 7b, 2023. 2

[10] Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. *arXiv preprint arXiv:1712.05558*, 2017. 2

[11] 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. Iglu 2022: Interactive grounded language understanding in a collaborative environment at neurips 2022, 2022. 2

[12] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 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, 2021. Association for Computational Linguistics. 2

[13] Vaibhav Kumar and Alan W Black. 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, 2020. Association for Computational Linguistics. 2

[14] Kaiqu Liang, Zixu Zhang, and Jaime Fernández Fisac. Introspective planning: Guiding language-enabled agents to refine their own uncertainty. *arXiv preprint arXiv:2402.06529*, 2024. 1

[15] Brielen Madureira and David Schlangen. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the codraw dataset, 2023. 2

[16] Brielen Madureira and David Schlangen. Taking action towards graceful interaction: The effects of performing actions on modelling policies for instruction clarification requests. *arXiv preprint arXiv:2401.17039*, 2024. 2

[17] Leora Morgenstern and Charles L. Ortiz. The winograd schema challenge: evaluating progress in commonsense reasoning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, page 4024–4025. AAAI Press, 2015. 2

[18] Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy, 2019. Association for Computational Linguistics. 2

[19] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2017–2025, 2022. 1, 2

[20] Jeongeun Park, Seungwon Lim, Joonhyung Lee, Sangbeom Park, Minsuk Chang, Youngjae Yu, and Sungjoon Choi. Clara: Classifying and disambiguating user commands for reliable interactive robotic agents, 2023. 2

[21] Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncer-

tainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023. 1, 2

[22] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005. 1