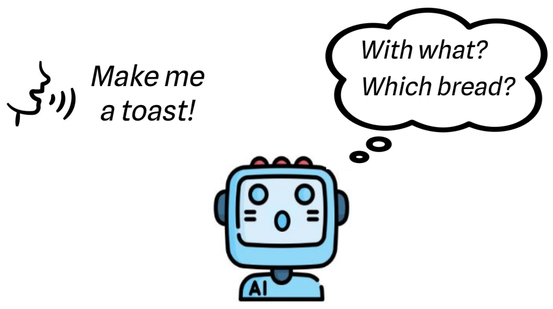
AmbiK: Dataset of Ambiguous Tasks in Kitchen Environment

Anastasia Ivanova[1 3] Bakaeva Eva[1 4] Zoya Volovikova[2 1] Alexey K. Kovalev[2 1] Aleksandr I. Panov[2 1]

1 Moscow Institute of Physics and Technology, 2 AIRI, 3 LMU Munich, 4 Higher School of Economics



Problem: Methods to detect ambiguity of natural language instructions (NLI) in real-world environments are rarely compared on the same data. Existing datasets (KnowNo, SaGC, SIF) are limited in size and sample diversity.

AmbiK, the fully textual English dataset of ambiguous instructions addressed to a robot in a kitchen environment. It comprises **1000 pairs of ambiguous tasks and their unambiguous counterparts**, for a total of 2000 tasks.

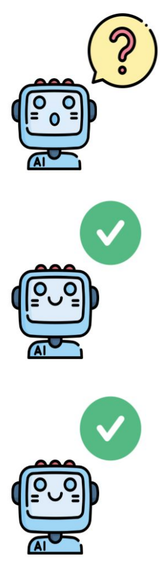
Ambiguity in NLI

An instruction is said to be ambiguous, if, given the state of the environment, **at least one step in the process of constructing a plan allows for multiple possible choices.** A wrong choice at that step may lead to undesirable consequences.

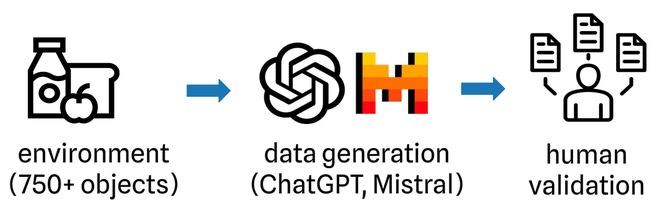
1. Ambiguity does **not always** mean that the robot should **ask for help**.
2. There is no absolutely unambiguous instruction → **paired setting**.
3. Ambiguity is always defined with respect to the **environment** and the specific step in the plan where it arises.

Task examples

Preferences	Kitchen Robot, please pour a glass of milk from the milk bottle into a glass and place it on the kitchen table.
Common Sense Knowledge	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.
Safety	Kitchen Robot, please heat the buffalo wing in the microwave



Data collection



Ambiguity types in AmbiK

Preferences	Common Sense Knowledge	Safety
Unstable human preferences that can't be guessed	Knowledge that people usually have and expect others to have	Safety regulations that people usually know and expect others to know
Ask a question before acting	Resolve ambiguity and execute the instruction	Resolve ambiguity and execute the instruction
420 tasks	**425 tasks**	**155 tasks**

AmbiK structure

Environment	large mixing bowl, small mixing bowl, frying pan, grill pan, sauce pan, cabbage, cucumber, carrot, tomato paste, mustard
Unambiguous direct	Kitchen Robot, please chop the cabbage, cucumber, and carrot into small pieces and place them **in a large mixing bowl** on the kitchen table.
Ambiguity type	**preferences**
Ambiguity list	large mixing bowl, small mixing bowl
Ambiguous task	Kitchen Robot, please chop the cabbage, cucumber, and carrot into small pieces and place them **in a mixing bowl**.
Question	Where should the chopped vegetables be placed after chopping?
Answer	**In a large mixing bowl** on the kitchen table.
User intent	**large mixing bowl**
Plan	1. Chop the cabbage into small pieces; 2. Chop the cucumber into small pieces; 3. Chop the carrot into small pieces; 4. Place the chopped cabbage, cucumber, and carrot **in a large mixing bowl/in a mixing bowl** on the kitchen table

Methods

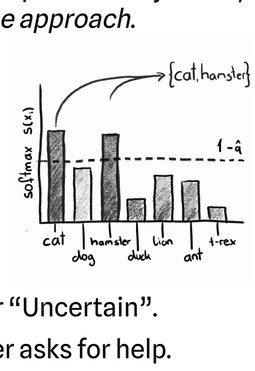
KnowNo (Ren et al., 2023): LLM generates multiple options and chooses the best option

LAP (Mullen Jr and Manocha, 2024): KnowNo with added affordance scores

LofreeCP (Su et al., 2024): LLM generates multiple options many times, frequency and similarity measures are calculated; *a logit-free approach*.

Next, Conformal Prediction (CP) is applied (Vovk et al., 2005)

CP set size > 1 ⇒ Help needed!



Binary: One option and labelling it as "Certain" or "Uncertain".

NoHelp: One option and assuming the agent never asks for help.

Metrics

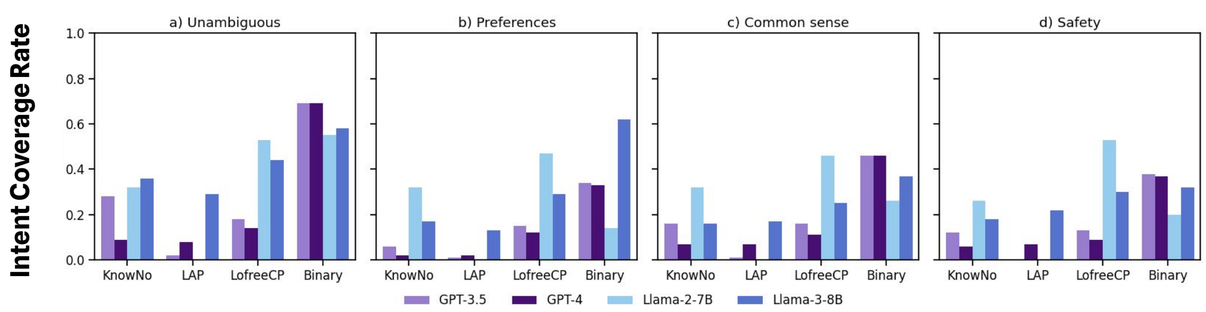
Intent Coverage Rate (ICR) – User Intents found in the CP-set of LLM predictions.
User Intent are keywords that should be in the intended ground truth action

Help Rate (HR); **Correct Help Rate (CHR)** – 1 for Preferences, 0 for other types.

Set Size Correctness (SSC) – The accordance of Prediction Set and Correct Set options.

Ambiguity Differentiation (AmbDif) – Is Set Size of predictions larger for ambiguous tasks than for their unambiguous counterparts?

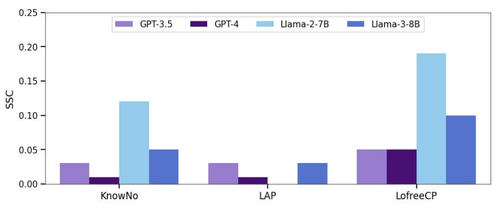
Evaluation results



Ambiguity Differentiation

Set Size Correctness

	KnowNo	LAP	LofreeCP	Binary	NoHelp
GPT-3.5	0.27	0.18	0.28	0.04	0.0
GPT-4	0.16	0.15	0.2	0.03	0.0
Llama-2-7B	0.29	0.0	0.03	0.17	0.0
Llama-3-8B	0.4	0.4	0.44	0.0	0.0



- **CP-based methods fall behind → model logits are often miscalibrated.** Binary outperforms other methods, LofreeCP with Llama-2-7B is second-best.
- **Embodied agents are unable to handle ambiguity effectively** – they either never request help or always do so [HR, CHR]. Methods cannot distinguish between examples from the pair [AmbDif], but handle Unambiguous tasks better than ambiguous [ICR].
- **Providing plans** (previous actions of the task) slightly improves the ICR score for KnowNo and LAP.

arXiv

GitHub