Open-Universe Assistance Games

Rachel Ma* MIT CSAIL
Jingyi Qu MIT CSAIL
Andreea Bobu MIT CSAIL
Dylan Hadfield-Menell MIT CSAIL

Abstract

Embodied AI agents must infer and act in an interpretable way on diverse human goals and preferences that are not predefined. To formalize this setting, we introduce Open-Universe Assistance Games (OU-AGs), a framework where the agent must reason over an unbounded and evolving space of possible goals. In this context, we introduce GOOD (GOals from Open-ended Dialogue), a data-efficient, online method that extracts goals in the form of natural language during an interaction with a human, and infers a distribution over natural language goals. GOOD prompts an LLM to simulate users with different complex intents, using its responses to perform probabilistic inference over candidate goals. This approach enables rich goal representations and uncertainty estimation without requiring large offline datasets. We evaluate GOOD in a text-based grocery shopping domain and in a text-operated simulated household robotics environment (AI2Thor), using synthetic user profiles. Our method outperforms a baseline without explicit goal tracking, as confirmed by both LLM-based and human evaluations.

1 Introduction

While AI agents can respond to diverse, complex human goals and preferences, their decision-making is often opaque and difficult to interpret. Our goal is to build AI agents that track and update explicit hypotheses of plausible user goals in open-ended interaction. Current interpretable AI agent designs depend on static sets of predefined goals. As a result, they struggle in open-universe environments where human preferences, goals, constraints, and other requirements are diverse, dynamic, and under-specified [4]. For example, a home grocery assistant may need to account for users with allergies, preferences for local ingredients, or specific dietary requirements. Even though this task is relatively well scoped, it is difficult for designers to anticipate this long tail of preferences in advance. While LLM agents that take actions based on full dialogue history can, in theory, adapt to a wide range of preference structures, their internal representations are hard to interpret. They also struggle to capture uncommon, user-specific, or novel requirements that are rarely represented in their training data.

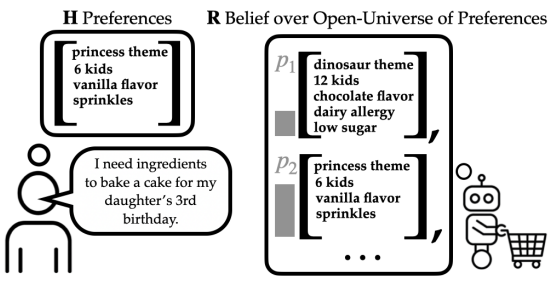

Figure 1: Our paper introduces *Open-Universe Assistance Games*, an assistance framework that models evolving user goals from an open-ended space of potential preferences. This framework reduces the specification effort for designers while supporting flexible, interpretable, and corrigible AI agents.

We build on Open-Universe Partially Observed Markov Decision Processes (OU-POMDPs) [31], which extend POMDPs to domains with unknown or changing sets of physical objects and relations.

*Corresponding Author (*Email*: rachelm8@mit.edu)

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

However, building human-centered agents also requires inferring goals and intent that are not predefined. Instead, these goals need to be identified from interaction, often from natural language. We formalize this problem as an Open-Universe Assistance Game (OU-AG). In an OU-AG, the human's preferences are modeled with an unknown number of latent dynamic preference structures. In this work, we will focus on *goals*, although the framework can model a wide range of potential preference structures.

To solve OU-AGs, we introduce **GOOD** (**GO**als for **O**pen-ended **D**ialogue). GOOD uses LLMs to (1) extract candidate goals from dialogue, (2) generate new goal hypotheses and prune unlikely ones, and (3) provide rankings over these goals. GOOD combines the adaptability of offline preference learning methods [21, 32] with the data efficiency of online learning [2, 13] while supporting natural and unconstrained conversations.

In this paper, we make the following contributions:

1. We introduce Open-Universe Assistance Games to model dynamic human preferences in open-ended natural language interaction;
2. We propose GOOD, a method for explicit goal hypothesis tracking and ranking from natural language dialogue;
3. We evaluate GOOD in two open-ended assistance domains: grocery shopping and home robot assistance. We show that GOOD substantially improves action quality in comparison to baselines that select actions based on full conversation context alone.

2 Background and Related Work

2.0.1 Interactive Agents with LLMs

LLMs are increasingly being used in few-shot task planning and for communicating between agents for large environments for embodied agents [11, 3, 12] as they carry extensive commonsense knowledge which can be helpful for better reasoning and decision making [36]. Recent approaches demonstrate improved instruction following capabilities [33, 18] however they typically rely on finetuning to constrained datasets and are limited to simple, structured instructions or closed-domain tasks. Our work is the first to infer rich, complex, and open-domain human preferences through natural language interaction in open-universe scenarios in an online method. Additionally, our modular architecture supports flexible integration with other action agents such as ReAct and RAISE after human goals are inferred. We compare our work directly to ablation baselines to show the significance of explicit goal tracking and a probabilistic inference method.

Previous works such as [23, 24, 10, 30] show that LLM agents can have different roles to achieve tasks. Our work leverages this concept and has different LLM calls that focus on specific tasks in our pipeline to be more efficient with how much information each LLM call has in memory. [23] also show that LLMs can be used to simulate human behaviors, something we take advantage of for our experiments in our work.

2.0.2 Preference Learning and NL Probabilistic Reasoning with LLMs

Offline preference tuning methods [21, 32] are data heavy but are generalizable to many domains and tasks. Online methods are data efficient but often task specific [2]. Our method bridges these by being both online and general-purpose. LLMs have recently been shown to be effective for supporting probabilistic reasoning in natural language [16, 1]. Prior approaches like [7, 9] rely on querying the human with best-of-k or structured queries and expect structured responses like binary comparisons [26, 34, 15]. In contrast, our work does not assume a specific interaction format and supports flexible dialogue, enabling the discovery and representation of complex goals that are not predefined from natural, open-ended human dialogue.

2.0.3 Evaluations/Benchmarks

While there are datasets and benchmarks that exist for multiple round dialogue such as [17, 27, 35], we want to be able to have dialogue in real time to various preferences and informed by actions and be able to take actions in response to the dialogue, and be able to evaluate the outcomes. Hence, we

design a synthetic conversation generator provided a human profile. We perform both LLM-as-a-judge [37] and human evaluations.

3 Open-Universe Assistance Games

In this section, we will introduce Open-Universe Assistance Games (OUAG). As an example, we model a grocery shopping agent that is gathering ingredients for a cake according to the human's preferences. Since the scenario involves reasoning about uncertainty, we model it as a POMDP, then extend to open-universe POMDPs, and connect to assistance game settings. This will help highlight how uncertainty over undefined human preferences/goals can be integrated into decision-making frameworks for interactive agents.

In general, the "preference structure" for the human agent may include explicit goals, constraints, or other forms of specification beyond traditional (static) reward functions. In this work, we will focus on "goals" as a practical instance of this broader class, but the framework naturally extends to richer and more complex specifications expressible in natural language.

3.1 Preliminaries

3.1.1 Partially-Observed Markov Decision Processes

We begin by formally introducing POMDPs and using that framework to model the grocery agent. In this case, we will model uncertainty about the store's inventory.

Formally, a *Partially Observed Markov Decision Process* (POMDP) is a tuple $\langle S, A, O, T, \Omega, r \rangle$, with the following definitions:

- S is a set of environment states;
- A is a set of actions that the agent can take;
- O is the set of observations, including the results of search queries in the inventory
- $T(s, a, s')$ is the transition model. It describes the probability distribution over the next state s', given the previous state and action[2];
- $\Omega(o_t \mid s_t, a_t)$ is the observation model. It defines a distribution over observations, given the previous state and action;
- $r(s)$ is a reward function that describes the agent's goal;
- $\gamma \in [0, 1)$ is the discount factor.

In a POMDP, the goal is to maximize the cumulative discounted reward $\mathbb{E}\left[\sum_t \gamma r(s_t)\right]$. A solution to a POMDP is a policy π that maps the action-observation history $\{(a_t, o_t)\}$ to a probability distribution over the current action $\pi : (A \times O)^* \to A$. A classic result states that optimal POMDP policies only depend on the agent's *belief* about the latent state. This allows us to abuse notation and write policies as functions of a distribution over states $\pi : \Delta(S) \to A$.

In our grocery shopping example, the state space has two components: 1) an inventory that tracks whether an item is in stock; and 2) a cart that tracks which items are queued for purchase. The observations are the success of adding an item to the cart and the results of search queries. The actions include searching the inventory, adding items to the cart, and checking out. The reward function indicates how well the cart matches the user's preferences and is zero until the items are purchased. The belief state is a distribution over which items are in stock. Optimal policies query the inventory to learn which items are in stock, add items to the cart, and balance the reward for the final cart against the total number of actions.

3.1.2 Open-Universe POMDPs

Representing our example as a POMDP requires that we pre-specify which items could be present in the store. In many cases, this will be challenging. The class of *Open-Universe* POMDPs [31] (OU-POMDP) addresses this shortcoming by modeling problems with an unknown number of objects.

[2]Note that we avoid modeling the distribution over the initial state and fold it into the transition distribution to reduce notation.

Formally, this involves modeling a set of object types. The state consists of a set of these types. Concretely, an *Open-Universe Partially-Observed Markov Decision Process* is a tuple: $\langle \{S, \Theta\}, A, O, T, \Omega, r, \gamma \rangle$.

The definitions are as before, except that the state space is now a tuple $\left(s, \{\theta_t^i\}\right)$of an environment state $s_t \in S$, as before, and a *set* of objects, each of type $\theta_t^i \in \Theta$. The transition function now maps over these tuples $T\left(\left(s_t, \{\theta_t^i\}\right), a_t, \left(s_{t+1}, \{\theta_{t+1}^i\}\right)\right)$.

In our grocery domain, S is the cart state, and the inventory state consists of an unknown number of items. A type is represented by an item description and whether or not it is in stock. The agent's belief now tracks three things: 1) how many items are in the inventory; 2) the description of the items; and 3) if it is in stock.

3.1.3 Assistance Games

POMDPs and OU-POMDPs account for uncertainty over the world state, but not over the task. In order to extend these models to account for uncertainty about goals, *assistance games* (AG) makes two changes to a POMDP formalism.

First, AGs model two actors: the human $\mathbf{H}$ and the robot $\mathbf{R}$. Second, AGs include a type $\theta \in \Theta$ for $\mathbf{H}$ that describes $\mathbf{H}$'s preferences. Only $\mathbf{H}$ observes θ. The robot $\mathbf{R}$ infers θ from $\mathbf{H}$'s actions.

Various forms of assistance games have been proposed [5, 8, 29]. We will build on *Cooperative Inverse Reinforcement Learning* (CIRL), which formalizes an assistance game with a fixed preference type and fully observed environment. Concretely, a CIRL game is a tuple: $\langle \{S, \Theta\}, \left\{A^{\mathbf{H}}, A^{\mathbf{R}}\right\}, T, r, \gamma \rangle$.

The overall state of a CIRL game is a tuple (s_t, θ) of environment state and preference type. A solution is a pair of policies $(\pi^{\mathbf{H}}, \pi^{\mathbf{R}})$ that specifies behavior for both actors. Both depend on the history of states and actions. The human policy additionally depends on the human's type. Note that the only uncertainty in a CIRL game is about the preference state. As our focus in this work is preference uncertainty, we will model our environment as fully observed. ([6] formalizes an AG with a partially observed environment state.) In this work, we will focus on assistance games where preference types θ are *goals* $g \in G$. Each g encodes a set of states $S_g \subset S$ that satisfy the goal. The reward function is 1 where the goal is satisfied and 0 elsewhere: $r(s, g) = 1$ **if** $s \in S_g$ **else** 0.

In our grocery shopping example, we make a few additions to model communication with the human. First, we can model $\mathbf{H}'s$ goal g as a desired shopping cart and define r accordingly. Next, we add a dialogue action, which asks $\mathbf{H}$ a question in natural language. Finally, $\mathbf{H}$'s actions $A^{\mathbf{H}}$ are defined to be natural language responses. These reveal information about g.

The robot's policy depends on the conversation history and the current cart. The human's policy depends on the conversation history, the current cart, and their desired cart. $\mathbf{R}$'s belief is a probability distribution over the possible desired carts G. The optimal policy for $\mathbf{R}$ will ask questions to reduce uncertainty about g and identify relevant items. This naturally trades off the cost of learning about the goal with the robot's improved decision quality.

3.2 Open-Universe Assistance Games

While assistance games model uncertainty about the goal, typical approaches to assistance games will rely on a designer to specify a set of possible goals in advance. One approach might be to use a large set of possible goals, such as a natural language representation of goals. In this case, $\mathbf{R}$'s belief tracks the most likely goal for $\mathbf{H}$.

In principle, this can represent any goal. In practice, the size of the goal space makes it difficult to track an explicit and interpretable belief over goals. To address this challenge, we adopt the same approach from OU-POMDPs. We model $\mathbf{H}$'s preferences as an evolving collection of latent preference objects (i.e., goals). This allows us to track an interpretable belief over $\mathbf{H}$'s active preferences.

We formalize this as a (dynamic) *Open-Universe Assistance Game*(OU-AG). An OU-AG is represented by a tuple: $\langle \{S, \Theta\}, \left\{A^{\mathbf{H}}, A^{\mathbf{R}}\right\}, T, r, \gamma \rangle$.

The key factor that differentiates an OU-AG from other AGs, such as CIRL, is that an OU-AG has states that consist of an environment state and an evolving set of *preference types*: $s_t, \{\theta_t^i\}$. The transition function also includes a distribution over the next preference set. Similarly, the reward

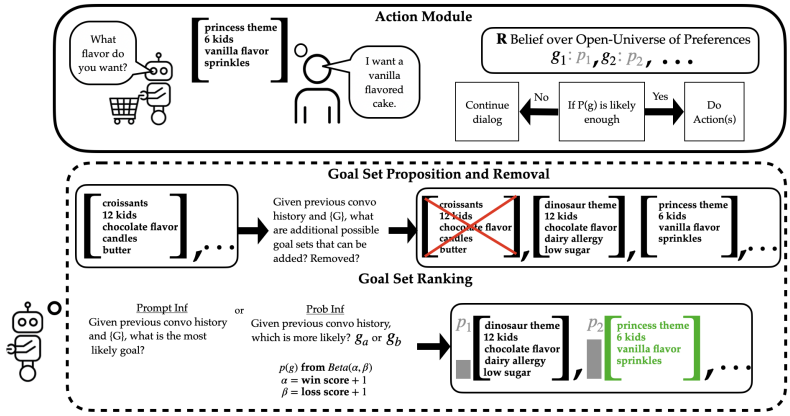

Figure 2: To solve Open-Universe Assistance Games, we propose the *GOals from Open-ended Dialogue* (GOOD) Architecture. This approach maintains hypotheses over user-goals with three modules: 1) a goal-proposal module that proposes new goals based on dialog; 2) removes goals that are no longer likely or relevant (potentially because they have been achieved); 3) prioritizes goals to guide action selection.

function r extends to depend on the full preference set $r(s, \{\theta_t^i\})$. As before, we let the preference types be goals, $g \in G$. The reward function is defined to be the number of active goals that are satisfied.

For example, consider an OU-AG formulation of our grocery domain. Initially, the human's goal is a single generic goal: "buy cake ingredients." After some rounds of dialogue, they refine their initial goal to "buy vanilla cake ingredients for 12" and add a goal "don't buy dairy" after being asked for additional considerations. Eventually, the set of goals might refine to reference specific brands of ingredients or account for what they have at home. During this interaction, $\mathbf{R}$ tracks a probability distribution over these goals and takes actions or asks questions accordingly. Once items are added to the cart, they stop being goals for $\mathbf{H}$. This representation allows us to track interpretable beliefs over goals throughout an interaction.

4 Goals from Open-Ended Dialogue

In this section, we present *GOals from Open-Ended Dialogue* **(GOOD)**, an agent design for solving Open-Universe Assistance Games. GOOD achieves three things: (1) proposing a finite list of goal sets that are plausible from the interaction with the human, (2) removing goals that are not relevant, and (3) ranking of these goals for an action agent or planner to act upon.

The key challenge that an OU-AG presents is that there is not a small set of latent states. To deal with this, we design inference with two parts. First, we track the plausible goals through proposing changes to goals, sampling new goals, and removing goals based on the conversation. Once we have a reasonable number of candidate goal sets, the *Inference* step determines a distribution over the candidates. Finally, the *Action* module takes actions to accomplish sufficiently likely goals.

To track plausible sets of goals we leverage a large language model to refine existing goals or propose new ones based on the last round of dialogue. It operates based on chunks of dialogue and updates as follows. First, it generates new candidate goal sets, based on the existing set of goals and the latest round of dialogue. Then, it ranks these hypotheses based on likelihood and removes the least likely sets of goals. You can see the prompt we use for generating new candidates in the Technical Appendix.

To rank the goals, we consider two designs. The first is simply to prompt an LLM to select the most likely from the list. Each goal set is prompted with a LLM to determine whether it should be removed. The most likely set of goals is given to the action module, especially in the o4-mini experiments.

Our second design is a more explicit inference module that attempts to compute a distribution over these sets of goals. We elicit pairwise comparisons from the LLM for which of two sets of goals is more likely, given the dialogue. Our choice of binary pairwise comparisons follows extensive prior work that shows LLM-based evaluation to be more stable and accurate with pairwise judgments

[25, 19]. Then, we track wins and losses for different sets of goals to assess the likelihood. We handle ties by letting the LLM classify both as likely (a win for both) or both as unlikely (a loss for both). For ties, a weight of 1 is added to the corresponding scores. If one is more likely than the other, a weight of 2 is added to the corresponding scores. We use a Beta distribution to model the 'true' win rate for a goal set. If α and β are the number of wins and losses (+1), then we remove goal sets based on the mean $\frac{\alpha}{\alpha+\beta}$. We pass goal sets to the action module if the mean is above a threshold and the variance is below a threshold.

The action module queries an LLM for the most appropriate action based on the belief about the human's goal set. To accomplish this, we prompt an LLM with the dialogue history and all sufficiently likely goal sets. This lets the action module prioritize from the set of available actions. In addition to taking actions in the world, it can also choose to continue the conversation.

This inference process is crucial to solving OU-AGs, by enabling the agent to maintain and refine beliefs over a dynamically constructed hypothesis space of goals.

5 Experiments

To show that having an agent with an explicit goal tracking system to infer over human preferences in Open-Universe Assistance Games settings, we run experiments in two domains to achieve diverse and complex human profiles: a grocery shopping agent domain and a household robot in AI2Thor simulation domain [14]. We compare GOOD to an ablation baseline that help emphasize the benefits of explicit goal tracking, and show that GOOD performs better than the *Full Context Baseline* agent which lacks explicit goal modeling and uses full conversation history as input for decision-making and action planning. We also show results of relying on different goal ranking methods for GOOD for goal inference: with simple prompting for the most likely goal and which goals to remove (*GOOD with prompt inference*), or with pairwise comparisons for generating a explicit distribution (*GOOD with probability inference*). Keep in mind that other agents can be easily substituted into GOOD to serve as action planners or for goal rankings during inference.

5.1 Domains

5.1.1 Grocery Domain

In the grocery shopping domain, the agent roleplays a shopping agent that is supposed to make purchases for the human. Their task is to identify a shopping basket that matching the human's preferences. The human roleplays given a lengthy human profile description that contains information about their preferences over ingredients for baking cakes (for five profiles) and making dinners (for the remaining five profiles). Our experiments feature 10 distinct human profiles for the grocery domain that covers various preference combinations over textures, flavors, allergies, and specific ingredients or inspirations, and varying levels of specificity for the final outcome for homemade cakes and dinners. See the Technical Appendix for the full human profile descriptions.

The possible actions that the agent can take for the Grocery Shopping experiment are `have dialogue` for n rounds (n defaulted to 2), `confirm basket`, `search inventory`, `add item to cart`, `remove item from cart`, and `buy basket and end task`. Inventory search on the Kaggle Grocery Store inventory dataset [28] relies on a semantic embedding search by similarity to narrow down options, and relies on another LLM call to retrieves a single most similar item. Cart manipulation functions are handled via dictionary operations. The main evaluation metric for the grocery domain is the *Cart score*, which is how well the outcome cart aligns with the human's task, preferences, and constraints given the description of the human profile. We do LLM-as-a-judge ratings and conduct Google surveys on Prolific [22] for human ratings. They were asked to provide a rating out of 0 to 10 (0 is completely unsuitable and 10 is perfectly aligned) while considering the preferences of flavors, textures, and lifestyle factors, and to be be strict about allergies and forbidden items. They are also asked to list issues and violations with the carts, along with a written justification for why they assigned the score they chose.

5.1.2 Robot Domain

In the AI2Thor robot domain, the robot is tasked with interacting with objects in the environment and the human to accomplish the human's preferences. Our experiments feature four human profiles that covers bringing or rearranging objects within the environment: two for bringing different food ingredients and kitchenware for breakfast and two for rearranging objects on the desk. See the Technical Appendix for the full human profile descriptions.

In the Home Robot domain (AI2Thor), available actions include physical manipulations such as `Open`, `Close`, `Pickup`, `Put`, `Toggle On/Off`, and domain-specific verbs like `Slice`, `Cook`, and `Clean`, as well as general actions like `Have Dialogue`, `Confirm Choices`, and `End Task`. Plans are executed sequentially. The main metric for the robot domain is the *Action score* based on five subcategories. There is a different rubrics designed that is specific for each human profile. To provide a brief summary: Preference Alignment checks if the actions performed followed preferences, Completeness checks for any missing critical actions or if the final outcome is present where the human wanted, Efficiency checks for redundant or unnecessary actions, Simplicity/Appropriateness/Safety checks for harmful actions, and Responsiveness to Feedback if the agent interacts with the human. The total score is 25. We do LLM-as-a-judge ratings and conduct Google surveys on Prolific [22] for human ratings.

5.2 Hyperparameters and Models

To simulate naturalistic dialogue readily for experiments, we use GPT-4o-mini [20] (with temperature 0 and top p 0.1) to generate both agent queries and human responses. For each round of dialogue, the agent generates a query based on its (1) high level task description, (2) the conversation transcript so far, (3) and a dialogue subtopic to inquire about. The human is modeled using a predefined lengthy *human profile* that encodes preferences relevant to a task. The human response is then generated using the human profile, the robot query, the subtopic, and the description of the current status of the task and environment. To estimate the likelihood for different goal sets, the Inference Module performs parallel batches of pairwise comparisons with an GPT LLM (we compare GPT-4.1-mini with temperature 0 and top p 0.1 or GPT-o4-mini with temperature 0). We sample 30% of the total number of possible goal set pairs. A goal set is considered sufficient to take action on if it exceeds the mean threshold 85% and if it is lower than the variance threshold 2%. Goal sets are removed if they have loss rates of 7 or more. These hyper-parameters are flexible and can be adjusted. For the action module, we compare using either GPT-4.1-mini and GPT-o4-mini as a planner through prompting. It is important to note that the module can be easily adapted to support any other planner, action agent, or language model.

5.2.1 Reproducibility

Despite setting the temperature to 0, GPT based models still contained variability in responses. To ensure reproducibility, we run six experiment trials for each human profile across both domains for GOOD and each baseline agent. An additional six trials per condition are conducted using GPTo4-mini to compare performances with a reasoning model. We evaluate experiments with LLM as a judge (GPT4.1-mini with temperature 0) and human evaluations. We conduct three independent LLM scoring runs per trial. Human evaluations were only done across three GPT4.1-mini trials due to cost and time. We report means and include standard error bars to reflect the variability across repeated runs.

5.3 Results

GOOD generally performs better than the *Full Context Baseline*, showing that an explicit goal module is helpful. This is still true when using GPT-o4-mini, a reasoning model. Keep in mind that the *GOOD with prompt inference* has a simpler inference method that relies on LLM prompting for the most likely and unlikely goals to act upon but *GOOD with probabilistic inference* features interpretable probabilistic quantities computed from pairwise goal set comparisons. In the robot domain experiments, it is clearer that GOOD excels by alternating dialogue with targeted actions. In contrast, *Full Context Baseline* often struggles after extreme amounts of repeated dialogue, and will repeatedly taking unhelpful or redundant actions, sometimes getting stuck. This is likely due to confusion during planning from the increasing dialogue context. By explicitly tracking goals,

Cart Scores and Time Performance (Grocery Domain)

Method	Cart Score (%)	Time (min)
Full Context (4.1-mini)	75.54 ± 0.85	3.86 ± 0.33
GOOD (prob inf, 4.1-mini)	80.42 ± 0.28	5.55 ± 0.52
GOOD (prompt inf, 4.1-mini)	84.07 ± 0.24	7.68 ± 0.89
Full Context (o4-mini)	78.30 ± 0.52	3.66 ± 0.26
GOOD (prob inf, o4-mini)	82.68 ± 0.24	10.20 ± 0.82
GOOD (prompt inf, o4-mini)	78.33 ± 1.01	43.05 ± 2.67

Table 1: Cart Scores and Time Performance (average mean ± SEM) for the Grocery domain. GOOD consistently outperforms *Full Context Baseline* that selects actions based on full conversation context. GOOD with probabilistic inference takes shorter to run on average than GOOD with simple prompt inference.

Action Scores and Time Performance (Robot Domain)

Method	Action Score (%)	Time (min)
Full Context (4.1-mini)	31.52 ± 4.12	4.10 ± 0.61
GOOD (prob inf, 4.1-mini)	66.44 ± 5.58	13.94 ± 2.20
GOOD (prompt inf, 4.1-mini)	53.96 ± 5.00	16.51 ± 1.95
Full Context (o4-mini)	44.64 ± 7.53	4.31 ± 0.94
GOOD (prob inf, o4-mini)	69.20 ± 4.29	19.45 ± 2.40
GOOD (prompt inf, o4-mini)	66.76 ± 5.21	54.20 ± 7.72

Table 2: Action Scores and Time Performance (average mean ± SEM) in the Robot domain. GOOD consistently outperforms *Full Context Baseline*. GOOD with probabilistic inference takes shorter to run on average than GOOD with simple prompt inference.

GOOD better focuses their actions to meet human preferences. GOOD with prompt inference takes the longest to run on average.

The scores are less differentiated in the Grocery domain compared to the Robot domain. The Robot domain contains more actions and different objects that lead to drastically different outcomes, making them more simple to evaluate and differentiate. The Grocery domain has fewer actions, and it is more difficult to compare and evaluate differences involving similar inventory objects (ie: a bag of lemons or a single lemon). Human evaluators on Prolific were given the same rubrics as the LLMs. In both domains, the LLM evaluations generally mirror the trends of the human ratings but differences are less pronounced. The Pearson Correlation coefficient between the human evaluations and the LLM-as-a-judge evaluations is 0.99 for the Grocery Domain and 0.85 for the Robot Domain. Results can be found in the Appendix: Table 3

6 Discussion and Conclusion

6.0.1 Limitations

This work focuses on text based scenarios. Incorporating a VLM (Vision Language Model) could enable for uncertainty and tracking goals from visual data as well. Human evaluations were conducted via Google Forms on Prolific, but the user experience could be improved with a more intuitive interface for presenting resulting text data. Human evaluations were not conducted across all trials and experiment types due to cost and time involved. Reproducibility with the o4-mini reasoning models is more difficult since temperature cannot be controlled. To mitigate this, we run multiple trials per method and human profile to report average performance. Because these human preferences are open-ended, tailored evaluation rubrics require effort to design. This further motivates why we perform human evaluations and include ratings about the general reasonableness of the outcome in all rubrics.

6.0.2 Conclusion and Future Work

We propose a framing of Open-Universe Assistance Games (OU-AG) to model interactions between humans and AI agents, enabling reasoning about human preferences. Unlike traditional cooperative games, we account for a unbounded space of possible complex natural language goals or preferences that are not predefined. We introduce a LLM assisted method, GOOD, to solve these OU-AG for open-ended goal inference that allows for explicit tracking of natural language goals and ensures that actions are only taken once the agent is certain enough about particular goals. We show that GOOD outperforms a baseline agent that only make action choices based on full conversation context, and that goals are useful to track to ensure human preference aligned behavior. Future work includes integrating GOOD and OU-AG with VLMs (Vision-Language Models) or other multimodal systems to support richer forms of input. Future work includes conducting more human subject studies to examine the benefits of interpretable goals, such as incorporating human feedback based on goals for corrections.

References

[1] David Eric Austin, Anton Korikov, Armin Toroghi, and Scott Sanner. Bayesian optimization with llm-based acquisition functions for natural language preference elicitation. *arXiv preprint arXiv:2405.00981*, 2024.

[2] Erdem Biyik and Dorsa Sadigh. Batch active preference-based learning of reward functions. In *Conference on robot learning*, pages 519–528. PMLR, 2018.

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

[4] Micah Carroll, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. Ai alignment with changing and influenceable reward functions. In *Proceedings of the 41st International Conference on Machine Learning*, pages 5706–5756, 2024.

[5] Alan Fern, Sriraam Natarajan, Kshitij Judah, and Prasad Tadepalli. A decision-theoretic model of assistance. *Journal of Artificial Intelligence Research*, 50:71–104, 2014.

[6] Andrew Garber, Rohan Subramani, Linus Luu, Mark Bedaywi, Stuart Russell, and Scott Emmons. The partially observable off-switch game. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27304–27311, 2025.

[7] Gabriel Grand, Valerio Pepe, Jacob Andreas, and Joshua B Tenenbaum. Loose lips sink ships: Asking questions in battleship with language-informed program sampling. *arXiv preprint arXiv:2402.19471*, 2024.

[8] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

[9] Kunal Handa, Yarin Gal, Ellie Pavlick, Noah Goodman, Jacob Andreas, Alex Tamkin, and Belinda Z Li. Bayesian preference elicitation with language models. *arXiv preprint arXiv:2403.05534*, 2024.

[10] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

[11] 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.

[12] 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.

[13] Ashesh Jain, Shikhar Sharma, Thorsten Joachims, and Ashutosh Saxena. Learning preferences for manipulation tasks from online coactive feedback. *The International Journal of Robotics Research*, 34(10):1296–1313, 2015.

[14] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[15] Volodymyr Kuleshov and Kevin Ellis. Active preference inference using language models and probabilistic reasoning. *arXiv preprint arXiv:2312.12009*, 2023.

[16] Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*, 2023.

[17] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

[18] Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models. *arXiv preprint arXiv:2401.02777*, 2024.

[19] Adian Liusie, Potsawee Manakul, and Mark JF Gales. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. *arXiv preprint arXiv:2307.07889*, 2023.

[20] OpenAI. Openai api. `https://openai.com/`, 2023.

[21] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[22] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of behavioral and experimental finance*, 17:22–27, 2018.

[23] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[24] Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

[25] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, et al. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*, 2023.

[26] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[27] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8689–8696, 2020.

[28] Maxim Sakhan. Canadian superstore grocery data, Jan 2023.

[29] Rohin Shah, Pedro Freire, Neel Alex, Rachel Freedman, Dmitrii Krasheninnikov, Lawrence Chan, Michael D Dennis, Pieter Abbeel, Anca Dragan, and Stuart Russell. Benefits of assistance over reward learning.

[30] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36, 2024.

[31] Siddharth Srivastava, Xiang Cheng, Stuart J Russell, and Avi Pfeffer. First-order open-universe pomdps: Formulation and algorithms. Technical report, 2013.

[32] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[33] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[34] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

[35] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *arXiv preprint arXiv:2007.12720*, 2020.

[36] Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36, 2024.

[37] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

7 Appendix

7.1 Pseudocode

Algorithm 1 GOOD: GOals from Open-ended Dialogue

Require: Initialize empty G for candidate sets of goal sets, empty transcript t

Initialize: inf_ranking $\leftarrow \{\}$, round $\leftarrow 0$
while task not complete **and** round $<$ max_rounds **do**
  $(a, t, \text{completed}) \leftarrow \text{Action}(\text{LLM}, \text{inf\_ranking})$
  $G \leftarrow \text{add\_goals}(G, t)$
  for each goal g in G **do**
    if inf_ranking[g] $>$ remove_criteria **then**
      remove(g)
    end if
  end for
  if last action was dialogue **then**
    $(\text{inf\_ranking}) \leftarrow \text{Inference\_Update}(\text{LLM}, G, t)$
  end if
  if task completed **then**
    break
  else
    round $\leftarrow$ round $+1$
  end if
end while=0

Algorithm 2 Probabilistic_Inference_Update Subroutine

Require: Language model LLM, candidate sets of goal sets G, transcript t
Initialize: certain_sets $\leftarrow \emptyset$, remainder_sets $\leftarrow \emptyset$
for sampled goal pairs (g_0, g_1) from G **do**
  result $\leftarrow$ LLM.prompt(g_0, g_1, t)
  Update certain_sets, remainder_sets, win_scores, loss_scores based on result
end for
for each goal g in G **do**
  $\alpha \leftarrow$ win_scores$[g] + 1$
  $\beta \leftarrow$ loss_scores$[g] + 1$
  if Beta(α, β).mean $\geq$ mean_thresh **and** var $\leq$ var_thresh **then**
    Append g to certain_sets
  else
    Append g to remainder_sets
  end if
end for
return certain_sets, remainder_sets, win_scores, loss_scores =0

7.2 Human Profiles

For the grocery domain, the ten human profiles that are tested on:

1. "that your name is Zoe and that you want to have ingredients to bake a cake. You are a marketing manager, you are a very busy person - juggling project deadlines and managing a team. You are allergic to nuts and avoids anything with almonds, hazelnuts, or peanuts. You love cakes with rich textures, like sponge cakes or chiffon cakes. You prefer light, airy cakes with a balance of sweetness—nothing overly sugary. Your go-to is a classic lemon drizzle cake with a hint of tangy frosting. You also like casual conversation, and behave like a normal human."

2. "that your name is Gavin and that you want to have ingredients to bake a cake. You are a Mechanical Engineer, you are extremely busy - long work hours and tight deadlines. You are not allergic to anything but prefers to avoid overly complex flavors. You like cakes that are simple but satisfying, such as a traditional chocolate cake with a thick layer of buttercream frosting. You love a rich, moist cake that isn't too sweet, and you enjoy cakes with a bit of crunch, like a cake topped with chopped chocolate or a sprinkle of cocoa nibs. You also like casual conversation, and behave like a normal human."

3. "that your name is Emily and that you want to have ingredients to bake a cake. You are a Freelance Writer, your schedule is flexible but often hectic, with multiple projects at once. You are allergic to dairy and you prefer vegan desserts. You love light, plant-based cakes made with ingredients like coconut milk or almond milk. You enjoy cakes with seasonal fruits like strawberries or peaches. Your favorite is a fluffy vegan carrot cake with a creamy cashew frosting. You also like casual conversation, and behave like a normal human."

4. "that your name is Lena and that you want to have ingredients to bake a cake. Your profession is a graphic designer, your schedule is moderate busy as you work a 9 to 5 but you often take on side projects. You are not allergic to anything but you love experimenting with unusual flavors in cakes. You enjoy cakes with unique combinations, such as matcha and vanilla or lavender and honey. Your favorite cake is a moist lavender cake with honey buttercream frosting, decorated with edible flowers for a visually stunning finish. You also like casual conversation, and behave like a normal human."

5. "that your name is Ben and that you want to have ingredients to bake a cake. Your profession is that you are a grad student who is very busy with classes and schoolwork. You are allergic to gluten but enjoys gluten-free cakes. You have a sweet tooth and loves indulgent cakes that are rich and decadent. Your favorite is a gluten-free chocolate lava cake, with molten chocolate oozing from the center. You prefer cakes with bold flavors, like dark chocolate or raspberry. You also like casual conversation, and behave like a normal human."

6. "that you are highly sensitive to textures and smells in food—nothing mushy, slimy, or strongly scented. You're looking to put together a plain, texture-safe dinner that feels predictable and gentle.
7. "that you're a disciplined athlete who tracks macros obsessively and avoids anything with sugar or fluff. Your goal is to shop for a high-protein, performance-focused dinner that supports muscle recovery You also like casual conversation, and behave like a normal human."
8. "that you prefer traditional brands and foods from the past, and you're skeptical of modern products or packaging. You want to cook a cozy, nostalgic dinner that feels like it came from a mid-century kitchen. You also like casual conversation, and behave like a normal human."
9. "that you're a sustainability-driven prepper who only buys local, low-waste, or shelf-stable foods. You're shopping for a dinner that reflects resilience and could work even in a self-sufficient off-grid setup. You also like casual conversation, and behave like a normal human."
10. "You make food choices based on tarot readings and symbolic meaning, guided by mood and intuition. Tonight, you're curating a spiritually resonant dinner that aligns with your emotional and cosmic themes. You also like casual conversation, and behave like a normal human."

7.3 Robot Domain

For the robot domain, the four human profiles that are tested on:

1. "you are someone usually like to start your day with something filling and warm for breakfast. You tend to include a few things on your plate, especially if you have a bit more time in the morning. Sometimes you enjoy freshly made items, and you like options you can assemble together, and place them on the countertop. You also like casual conversation, and behave like a normal human.",
2. "You are someone who doesn't really spend much time on breakfast. Most days you just grab something quick—sometimes just a drink, maybe a small snack if you feel like it. You don't like a lot of fuss or cleanup in the morning. You also like casual conversation, and behave like a normal human.",
3. "You are someone who likes their workspace to be tidy and everything to have its place. You prefer to keep your laptop, pens, and books neatly arranged on your desk so you can easily find what you need. Clutter distracts you. You want help to arrange the objects in your room and on your desk. You also like casual conversation, and behave like a normal human.",
4. "You are someone who feels most comfortable when your things are spread out around you. Having objects within reach and a bit of creative mess inspires you. You aren't too concerned if your desk gets a little cluttered—it helps you feel at home and can even spark new ideas. You want help to arrange the objects in your room and on your desk. You also like casual conversation, and behave like a normal human.",

7.4 Robot Domain Implementation

If any action fails due to simple environment failures, the agent "undoes" prior actions by resetting the environment and replaying all actions from a successful action history. This undo mechanism is implemented manually since AI2Thor lacks native undo support. The robot uses teleportation to move between interactable object positions. Additional support logic ensures receptacles are opened as needed before executing `Pickup` or `Put` actions, and handles object pairings (e.g., stove burners and knobs) for tasks like toggling appliances.

7.5 Evaluations

Action Score Rubric (An example with Robot Domain Profile 1's Scenario)

For each checklist item:

- Clearly state the checklist item.
- Indicate whether the criterion was met (Yes), not met (No), or partially met (Partial).
- Provide a detailed explanation for your assessment, referencing specific actions from the transcript.

After evaluating all checklist items, for each of the five main categories below:

- Assign a score from **0 to 5**.
- Clearly explain the reasoning behind the score, referencing your earlier checklist assessments and the agent's actions.

Then, provide:

- The **overall score**, which is the sum of the five category scores (maximum 25), and show the calculation.
- **Additional comments**, including any insights, highlights, or concerns about the agent's behavior.

Human Profile: `{human_profile}`

Agent Action Transcript: `{action_list}`

1. **Preference Alignment**
   - Does the agent prepare a warm, filling breakfast (e.g., includes cooked eggs, toasted bread, or similar items)?
   - Are multiple breakfast items included, allowing for variety and assembly (e.g., eggs, toast, fruit/veggies, hot beverage)?
   - Are freshly made or cooked items prepared (not just pre-packaged or cold)?
   - Are items placed on the countertop or plate for the user to assemble?
   - Is casual conversation or confirmation included in the interaction?
2. **Completeness**
   - Are all essential breakfast components present (protein, bread/grain, beverage)?
   - Are any critical steps missing that would prevent the user from enjoying a full breakfast?
   - Are utensils and serving items provided as needed?
   - Are any nice-to-have items included (e.g., fruit, veggies, buttered toast)?
3. **Efficiency**
   - Are there no redundant or unnecessary actions?
   - Are irrelevant or unrelated actions avoided?
   - Is the number of steps reasonable and focused on the breakfast goal?
   - Are substitutions or alternatives reasonable and efficient?
4. **Safety and Appropriateness**
   - Are all food items handled safely (e.g., cooked properly, no cross-contamination)?
   - Are no harmful or inappropriate actions performed (e.g., using dirty utensils, unsafe appliance use)?
   - Are any dietary restrictions or allergies mentioned in the profile respected?
5. **Responsiveness to Feedback**
   - If the user expresses a preference or gives feedback, does the agent adjust actions accordingly?
   - Does the agent confirm or acknowledge feedback before proceeding?
   - Is the agent attentive to the user's needs and requests throughout the process?

Conversation Score Rubric (An example with Robot Domain Profile 1's Scenario)

You are an evaluator agent. Carefully evaluate the following transcript of a breakfast preparation interaction using the detailed rubric below.

For each subcategory, do the following:

- State the subcategory name.
- Clearly show the score awarded (0–5).
- Provide a detailed explanation justifying the score, referencing specific parts of the transcript.

After all subcategories, provide:

- The total score out of 50, showing the sum of all subcategory scores, with the calculation clearly shown.
- A brief summary of the overall evaluation.

Rubric

1. Information Gathering Effectiveness

- **1.1 Depth of Understanding**
  - 5: Thorough understanding of preferences (filling, warm, variety, freshly made, likes assembling, casual conversation).
  - 4: Understands most preferences; minor details missed.
  - 3: General understanding; lacks depth or misses important points.
  - 2: Limited understanding; surface-level only.
  - 1: Barely understands preferences.
  - 0: No understanding of preferences.
- **1.2 Breadth of Information**
  - 5: Explores multiple aspects (temperature, variety, assembly, timing, conversation).
  - 4: Covers most aspects; minor areas missed.
  - 3: Covers some aspects; several important ones left out.
  - 2: Narrow focus; very few aspects.
  - 1: Barely explores relevant aspects.
  - 0: No exploration.
- **1.3 Use of Dialogue to Learn More**
  - 5: Uses open-ended questions, follow-ups, clarifications to deepen understanding.
  - 4: Some follow-ups and clarifications; not very probing.
  - 3: Occasionally asks questions; relies mostly on initial info.
  - 2: Rarely asks questions or clarifications.
  - 1: Only yes/no or closed questions; no follow-ups.
  - 0: No engagement in dialogue.

2. Profile Representation Accuracy

- **2.1 Human Behavior Consistency**
  - 5: Consistently aligns with profile preferences.
  - 4: Mostly aligns; some vagueness.
  - 3: Some inconsistencies.
  - 2: Rare alignment.
  - 1: Contradicts profile.
  - 0: No alignment with profile.

- **2.2 Naturalness of Conversation**
  - 5: Casual, natural tone.
  - 4: Mostly natural; minor robotic moments.
  - 3: Some awkwardness; generally understandable.
  - 2: Frequently stilted.
  - 1: Very robotic or scripted.
  - 0: Incoherent.

3. Outcome Quality

- **3.1 Clarity of Breakfast Goals**
  - 5: Very clear goals (specific foods, preparation, assembly).
  - 4: Mostly clear; some ambiguity.
  - 3: Somewhat clear; lacks specificity.
  - 2: Vague or incomplete.
  - 1: Barely stated or confusing.
  - 0: No clear goals.
- **3.2 Agent's Appropriateness of Actions**
  - 5: Perfectly aligned with conversation flow.
  - 4: Mostly appropriate; minor missteps.
  - 3: Sometimes inappropriate actions.
  - 2: Frequently inappropriate.
  - 1: Rarely appropriate.
  - 0: Completely disruptive.

4. Overall Interaction Quality

- **4.1 Engagement Level**
  - 5: Engaging with positive tone.
  - 4: Mostly engaging; minor dullness.
  - 3: Somewhat flat or repetitive.
  - 2: Low engagement.
  - 1: Very low; frustration evident.
  - 0: No engagement; abandoned.
- **4.2 Coherence and Flow**
  - 5: Natural progression, smooth transitions.
  - 4: Mostly coherent; minor awkwardness.
  - 3: Somewhat disjointed but understandable.
  - 2: Frequently confusing.
  - 1: Very fragmented.
  - 0: Chaotic or nonsensical.

Human Profile: `{human_profile}`
Transcript: `{convo_transcript}`

Return your answer in this format:

1. For each subcategory:
   - Subcategory name
   - Score awarded / 5
   - Detailed explanation with transcript references
2. Brief summary of the overall evaluation
3. Final total score (out of 50), with calculation shown

Cart Rubric (Grocery Domain)

You are an evaluator agent reviewing a shopping cart based on a specific human profile and task. Carefully analyze whether the contents of the provided cart align with the following human profile and goals:

- **Human Profile:** `{human_profile}`
- **Cart to Evaluate:** `{cart}`

Your job is to:

1. Evaluate how well the cart aligns with the human's task, preferences, and constraints.
2. Identify any violations or issues (e.g., allergens, missing key ingredients, conflicting items).
3. Provide a rating score from 0 to 10 representing the overall suitability of the cart for helping the human achieve their goals while respecting their preferences and constraints.
   - 0 means completely unsuitable.
   - 10 means perfectly aligned and ideal.
4. Explain the reasoning behind your rating clearly and in a human-readable way.

Be strict about any allergies or forbidden items. Consider preferences on flavors, textures, and lifestyle factors.

Format your response like this:

- `cart_fit_rating: <integer 0-10>`
- `issues_found: [<list of violations or concerns, if any>]`
- `explanation: "<clear, human-readable explanation of how well the cart fits the human profile and task>"`

Conversation Rubric (Grocery Domain)

You are an evaluator reviewing a conversation transcript with respect to a human profile.

Given the human profile below and the conversation transcript, rate the overall quality of the conversation on a scale from 0 to 10, where:

- **0** = Completely poor conversation; no alignment with the human's preferences, constraints, or goals.
- **10** = Excellent conversation; fully aligns with the human's preferences and constraints, is natural and engaging, and effectively supports the human's goals.

Consider these factors:

- Understanding and respecting the human's preferences and constraints.
- Naturalness and engagement of the conversation.
- Clarity and support for the human's goals.
- Tone appropriateness and human-like behavior.
- Presence or absence of major issues or misalignments.

Return your answer in this format:

- `conversation_rating: <integer from 0 to 10>`
- `explanation: "<clear, concise justification citing specific strengths or weaknesses in the conversation>"`

Human Profile: `{human_profile}`
Conversation Transcript: `{convo_transcript}`

Human vs LLM-as-a-Judge Performance for Action/Cart Scores

Method	Human (%)	LLM (%)
Full Context (Grocery)	65.10 ± 2.34	76.21 ± 1.19
GOOD (prob inf, Grocery)	74.79 ± 2.66	81.58 ± 0.37
GOOD (prompt inf, Grocery)	76.77 ± 1.74	83.76 ± 0.42
Full Context (Robot)	43.80 ± 1.68	29.13 ± 2.54
GOOD (prob inf, Robot)	63.49 ± 1.91	75.93 ± 2.16
GOOD (prompt inf, Robot)	61.86 ± 1.77	48.13 ± 1.53

Table 3: Human evaluations compared to LLM-as-a-judge evaluations (average mean ± SEM) for the Grocery domain (Cart Score) and the Robot domain (Action Score). GOOD consistently outperforms *Full Context Baseline*. These are performed on three trials of 4.1-mini experiments for each method. Pearson Correlation for Grocery Domain is 0.99 and 0.85 for Robot Doman.

Conversation Scores (Robot and Grocery Domains)

Method	Domain	Convo Score (%)
Full Context (4.1mini)	Robot	76.04 ± 4.90
GOOD (prob inf, 4.1mini)	Robot	88.74 ± 0.56
GOOD (prompt inf, 4.1mini)	Robot	87.34 ± 0.52
Full Context (o4mini)	Robot	83.32 ± 3.53
GOOD (prob inf, o4mini)	Robot	87.98 ± 1.15
GOOD (prompt inf, o4mini)	Robot	85.00 ± 1.96
Full Context (4.1mini)	Grocery	73.89 ± 0.67
GOOD (prob inf, 4.1mini)	Grocery	76.61 ± 0.64
GOOD (prompt inf, 4.1mini)	Grocery	79.65 ± 0.29
Full Context (o4mini)	Grocery	78.53 ± 0.50
GOOD (prob inf, o4mini)	Grocery	77.36 ± 0.50
GOOD (prompt inf, o4mini)	Grocery	86.39 ± 0.26

Table 4: Conversation scores (mean ± SEM) for each method and model version in the Robot and Grocery domains.

7.6 Extra Results

Conversation Scores: Results for conversation scores can be found in Tables 4 and 5. GOOD generally outperforms the *Full Context Baseline* for conversation scores. GOOD with probabilistic inference performs comparably to GOOD with prompt inference in the robot domain. In the grocery domain, the conversation scores are quite similar across all methods. Human evaluators and LLM-as-a-judge trends mirror each other. Human evaluators generally remarked on *Full Context Baseline's* transcripts being extremely repetitive, and GOOD with probabilistic inference transcripts being clear, concise, and short in their comments.

8 Prompts

Begin on the next page due to Latex formats.

Human and LLM Conversation Scores (Robot and Grocery Domains)

Method	Domain	Human (%)	LLM (%)
Full Context	Robot	66.54 ± 1.50	84.63 ± 1.27
GOOD (prob inf)	Robot	87.22 ± 1.18	89.10 ± 0.77
GOOD (prompt inf)	Robot	88.57 ± 1.10	86.90 ± 0.72
Full Context	Grocery	68.55 ± 1.56	76.94 ± 0.39
GOOD (prob inf)	Grocery	75.78 ± 2.39	79.44 ± 0.61
GOOD (prompt inf)	Grocery	72.39 ± 1.42	79.86 ± 0.38

Table 5: Human and LLM conversation scores (mean ± SEM) for the Robot and Grocery domains. Human evaluations mirror the same rankings as LLM evaluations (performed on only three GPT-4.1-mini runs)

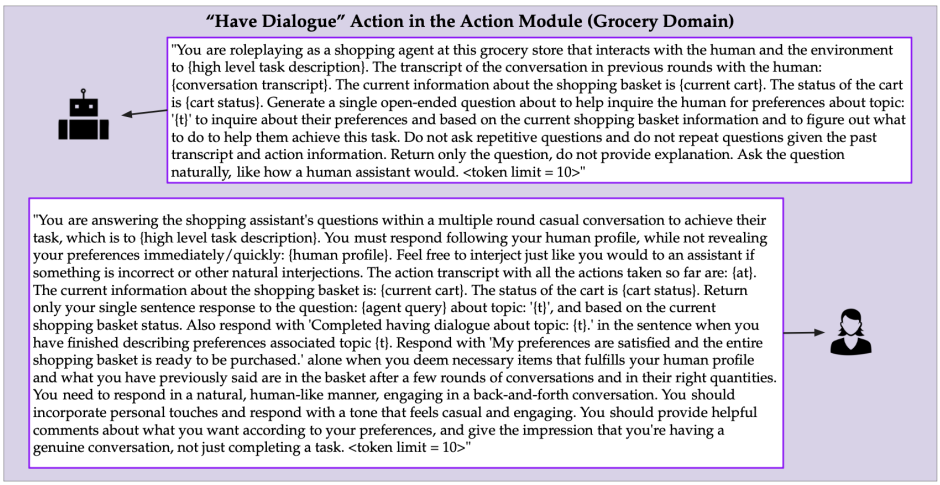

Figure 3: Agent query generation and human response generation prompts for the grocery domain.

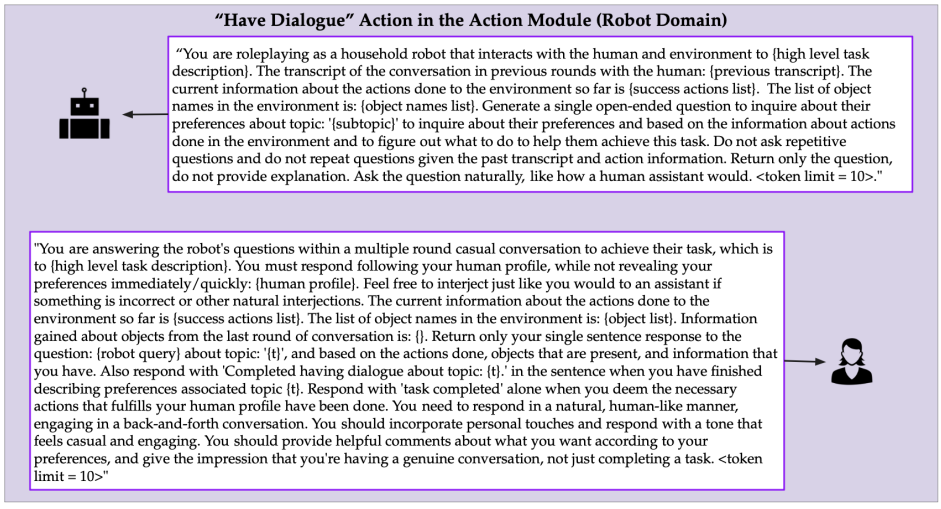

Figure 4: Robot query generation and human response generation prompts for the robot domain.

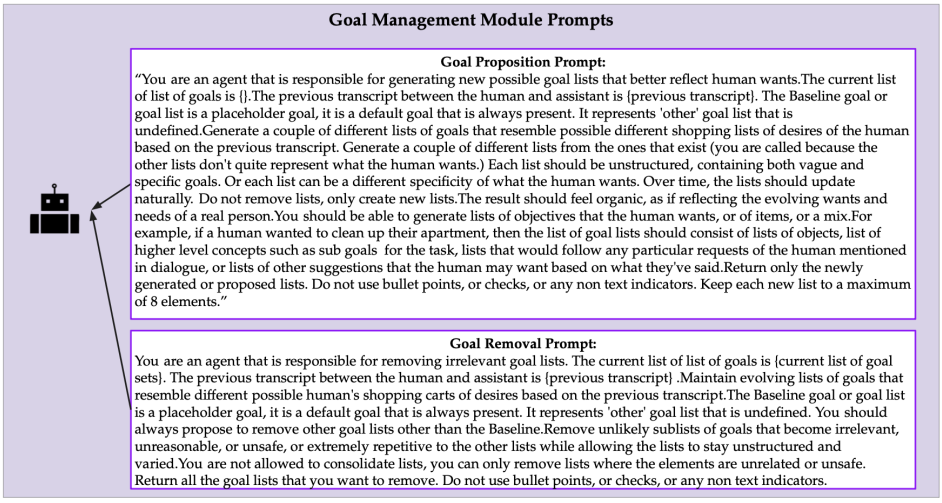

Figure 5: Goal Proposition Prompt and Goal Removal Prompt for GOOD.

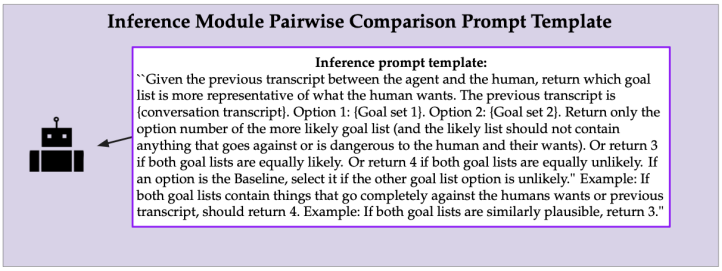

Figure 6: Inference Module Pairwise Comparison Prompt Template for GOOD.

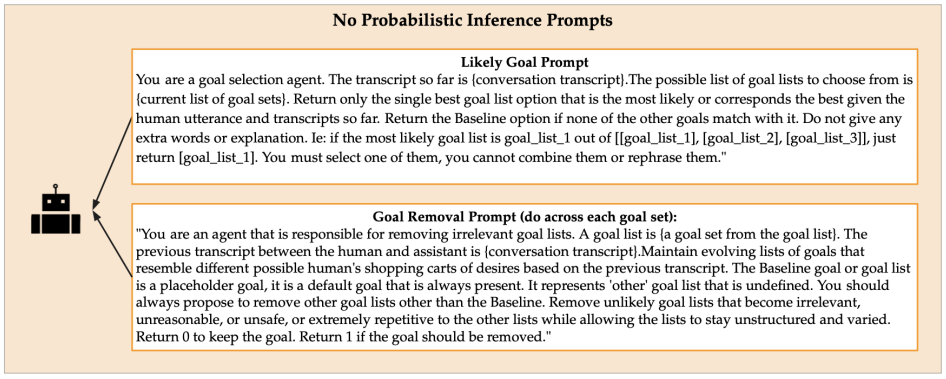

Figure 7: Prompting for the most likely goal and which goals to remove for the *No Probabilistic Inference Baseline*.

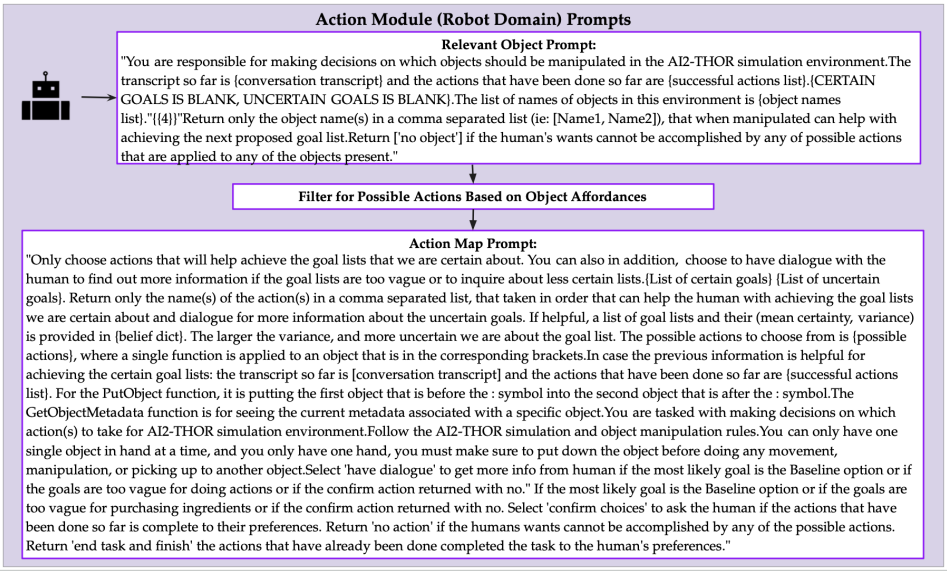

Figure 8: Action module prompts for the robot domain.

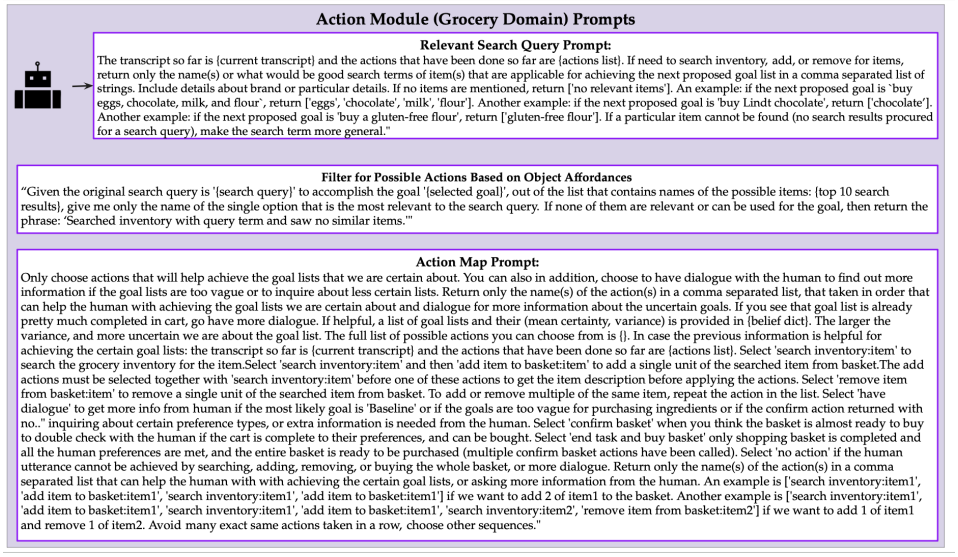

Figure 9: Action module prompts for the grocery domain.