Probing the Multi-turn Planning Capabilities of LLMs via 20 Question Games

Yizhe Zhang, Jiarui Lu, Navdeep Jaitly
Apple
{yizzhang, jiarui_lu, njaitly}@apple.com

Abstract

Large language models (LLMs) are effective at answering questions that are clearly asked. However, when faced with ambiguous queries they can act unpredictably and produce incorrect outputs. This underscores the need for the development of intelligent agents capable of asking clarification questions to resolve ambiguities effectively. This capability requires complex understanding, state tracking, reasoning and planning over multiple conversational turns. However, directly measuring this can be challenging. In this paper, we offer a surrogate problem which assesses an LLMs's capability to deduce an entity unknown to itself, but revealed to a judge, by asking the judge a series of queries. This *entity-deducing game* can serve as an evaluation framework to probe the conversational reasoning and planning capabilities of language models. We systematically evaluate various LLMs and discover significant differences in their performance on this task. We find that strong LLMs like GPT-4 outperform human players by a large margin. We further employ Behavior Cloning (BC) to examine whether a weaker model is capable of imitating a stronger model and generalizing to data or domains, using only the demonstrations from a stronger model. We finally propose to use Reinforcement Learning to enhance reasoning and planning capacity of Vicuna models through episodes of game playing, which lead to significant performance improvement. We hope that this problem offers insights into how autonomous agents could be trained to behave more intelligently in ambiguous circumstances.

1 Introduction

In uncertain circumstances, conversational agents may need to take the initiative to reduce their uncertainty by asking good questions proactively, thereby solving problems more effectively. This requires intricate, interactive, strategic decision-making and reasoning about the agent's next move in a multi-turn conversation. This capability is crucial in various applications like task-oriented chatbots, recommendations, and conversational search.

Traditional dialogue systems have been achieved by modularizing various aspects of such tasks into sub-tasks such as *understanding*, *state tracking*, *planning*, and *generation*. However, recent advances in LLM have made it possible to create an end-to-end pipeline for developing autonomous agents that can complete complex tasks using enhanced planning and memory capabilities. Promising autonomous agents, such as ReAct (Yao et al., 2022), AutoGPT (Significant Gravitas, 2023), and LangChain (Langchain-AI, 2023), require the underlying LLM to recall information from previous dialogues, resembling the *understanding* and *state tracking* stage. They also rely on the LLM to decompose larger tasks into more manageable components, analogous to the *planning* stage. Among them, some approaches (*e.g.*, HuggingGPT (Shen et al., 2023)) use a *static* planning strategy by first generating the complete plan via LLM and subsequently tackling subtasks. Other approaches (*e.g.*, AutoGPT) adopt an *interactive* planning strategy, where the generation of each action is conditioned on the outcome of the previous planning steps.

While LLM-powered agents are powerful, they can perform poorly when a user's instructions are unclear. This poses a critical challenge – how to accurately elicit and capture the user's intents, which are often nuanced and noisy, to enable dynamic rather than static human-computer interactions. For example, in Figure 1, the agent aims to: 1) accurately assess the current dialog state; 2) disambiguate user's intent and satisfy the user demand; 3) ask as few questions as possible.

Progress in this direction is difficult because directly measuring complex understanding, reasoning and planning capabilities is challenging. In this study, we investigate this somewhat overlooked research problem – *how good the LLMs are at asking*

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1495–1516

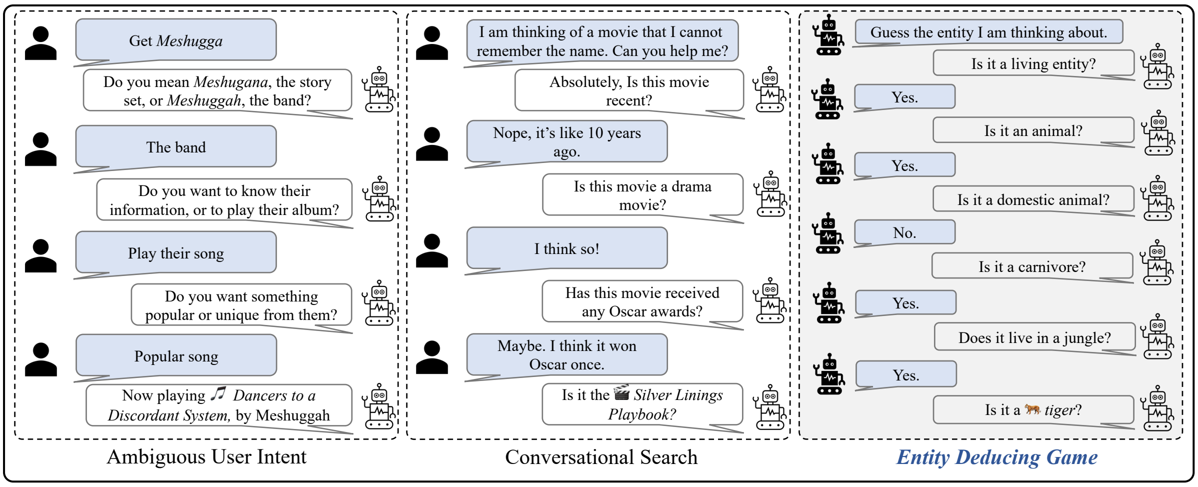

Figure 1: The entity deducing game resembles real scenarios where the agent may need to strategically ask clarification question based on the current conversation to elicit the actual user intent in as few turns as possible.

questions and deducing intent. Instead of directly evaluating disambiguation ability of LLMs, we study a *proxy task in a controlled setting with clear evaluation*. We propose to use entity-deducing games, specifically the 20 questions game (Q20) (Akinator, 2007), to assess the complex reasoning and strategic planning capability of LLMs, in formulating precise questions/guesses over long conversations (Figure 1). This serves as a first-step towards probing the underlying cognitive abilities required for effective disambiguation. This game requires a model to infer an unidentified entity with as few queries as possible. To achieve this, the model needs track the dialogue state over turns, and use its reasoning and planning skills to effectively partition and narrow down the search scope.

We systematically evaluated several LLMs on this task, and found significant differences in their conversational reasoning and planning capabilities. We then investigated a set of research questions to enhance open-source models using demonstrations from high-performing closed-source models. Finally, we employ PPO (Schulman et al., 2017) to improve vanilla Vicuna (Chiang et al., 2023) models using multi-turn game playing directly. We will provide the code and dataset to facilitate future research. Our findings offer insights into potential future advancements in autonomous agents that proactively lead the conversation.

2 Entity-deduction Arena (EDA)

During an entity-deducing game session, two players engage in a game regarding a specific entity. One player, "*judge*" (J), is provided with the entity and is expected to respond to queries or guesses from the guesser using only the responses "Yes," "No," or "Maybe,". The other player, "*guesser*" (G), is agnostic about the entity, and is supposed to pose a series of questions to deduce the entity using as few queries as possible.

The judge does not require access to the dialogue history and only needs the entity and current question to provide a response[1]. This task is akin to closed-book QA (Roberts et al., 2020), which current LLMs can reasonably handle. On the other hand, playing the guesser is more demanding. A proficient G necessitates several multi-turn dialogue capabilities working in synergy: 1) **State Tracking and Understanding**: G must comprehend multi-turn context, track asked questions, and understand its position in the game and coreference resolution. 2) **Strategic Planning**: G needs to strategically ask questions to progress efficiently towards a better state, avoiding redundant queries and ensuring consistency with prior knowledge. 3) **Inductive Reasoning**: G must use conversation comprehension to generate conjectures based on acquired knowledge. G must inherently establish a *taxonomy representation* to efficiently and accurately identify the correct entity among numerous options.

Our assessment, referred to as the *Entity-Deduction Arena* (EDA), focuses on evaluating the capability of various LLMs as the guesser, as a proxy to probe their overall capabilities in handling complex multi-turn conversational tasks involving

[1]Our experiment indicates that incorporating entire dialogue history negatively impacts accuracy as the additional information tends to confuse the judge.

proactively asking clarification questions.

While EDA may not evaluate all aspects of reasoning and planning ability, it capture several key capabilities which are important for any conversational AI agent that needs to actively elicit information from users through clarifying questions. The task provides a targeted way to probe these capabilities in a controlled setting, and provides a new and complementary benchmark compared to existing tasks for probing language models' capabilities from a different angle focused on multi-turn strategic conversation.

2.1 Experimental settings

Datasets We conducted the evaluation on two proposed datasets: *Things* and *Celebrities*. The *Things* dataset consists of 500 entities that are commonly found in the web corpus (Raffel et al., 2020). It encompasses a wide range of categories, such as common-life objects, animals, foods, plants, vehicles, clothing, professions, materials, instruments, places, birds, sports, buildings, furniture, celestial bodies, mythical creatures, events, activities. The larger category "objects" account for 20% percentage of the items. The *Celebrities* dataset consists of 500 celebrity names covers 32 different nationality, diverse eras of life and various occupations. Among the nationalities, the United States holds the largest proportion with 27.7%. In terms of occupations, the largest categories are actor/actress, accounting for 18.4%. Both datasets were divided into training, evaluation, and testing, with sizes of 300, 100, and 100 respectively. The composition of each dataset is provided in Appendix A.

Judge (`J`) We employ GPT-3.5-turbo as the judge. The judge takes the entity, questions from the guesser and the following prompt to generate a response of "Yes," "No," or "Maybe" for Things dataset. In guessing the celebrity name, the choices are "Yes," "No," or "Dunno". Consequently, the resulting prompt is slightly different. The prompt, and rationale for using "Dunno" for the *Celebrities* dataset are provided in Appendix B)

> *Based on your knowledge about the entity: {entity}, respond to the following question or guess. Limit your respond to only "Yes.", "No." or "Maybe.", with no explanation or other words. Never say the answer in your response. If the question is to solicit the answer, respond "No.". \n Question/Guess: {question} (Yes/No/Maybe)*

Whenever the correct answer is contained in the generation from `G` as an exact substring match, we manually set the output of `J` to be *"Bingo!"* and `G` wins this game. At the penultimate step of `J`, an extra prompt *"You must guess now, what's it?"* will be appended to `J`'s response to guide `G` in making the ultimate guess. To emulate more deterministic responses from `J`, we use a temperature of 0.2 for the generation. Admittedly, we observe that the judge model exhibits occasional inaccuracies in its responses, resulting in a noisy environment. We asked internal non-paid volunteers to annotate 300 responses from `J` for each dataset, and the error rate is 3.07% and 2.95% for *Things* and *Celebrities*, respectively. Nevertheless, this noisy environment of `J` is consistent with all models. Addressing judge inaccuracies and inconsistency is an interesting direction for future work.

Guesser (`G`) The guesser model receives the current dialogue history as input and generates the next question or final guess, guided by the instructions provided in the following (the prompt for *Celebrities* is provide in Appendix B):

> *Your task is to ask a series of questions to deduce the entity that I'm thinking of with as few queries as possible. Only ask questions that can be answered by "Yes," "No," or "Maybe". Do not ask for hint. Make your question brief with no linebreaker. Now start asking a question. \n {dialog history}*

We used a sampling approach with a temperature of 0.8 for all experiments, which sufficiently yields diverse outcome. The conversation format for each LLMs is based on their official guidelines.

Evaluation metrics We assess the model's performance by evaluating its final prediction using the Exact Match (EM) criteria [2]. This evaluation considers four key metrics: 1) **#Turns**, which represents the average number of turns taken to complete the game. Games terminate at 20 turns if failed. Lower is better. 2) **Success** rate, which indicates the percentage of games won by `G`. Higher is better. 3) **#Yes**, which represents the average number of "yes" responses received from the `J`. 4) **Score**, which is a combined game score of **#Turns** and success rate, defined in Eq. (1).

$$S = (1 - \lambda \cdot \max(\textbf{\#Turns} - 5, 0)) \cdot \mathbf{I}(\texttt{G} \text{ wins}) \quad (1)$$

[2]A more lenient evaluation metric could be employed, *e.g.* embedding similarities or LLM judges. However, these metrics rely on the specific model or LLM judge employed thus are less consistent than the EM.

where $\mathbf{I}(\cdot)$ represents the identity function, $\lambda = 0.02$ is a balancing hyperparameter. A higher value for λ would result in assigning inadequate rewards for successful games that take longer, while a smaller value would penalize the model too little for taking a longer turns. In our experiments, we discovered that a value of 0.02 led to models with a reasonable and stable reward. It is worth mentioning that reward design is more of a heuristic, and there may be superior choices available. The **#Yes** is more of a statistic than a evaluation metric, but we have observed some correlation between **#Yes** and the outcome. Intuitively, a losing game is often characterized by a high frequency of unproductive guesses (with "No" or "Maybe" response from J).

3 Benchmarking LLMs on EDA

We assess several widely-used LLMs, such as GPT-4, GPT-3.5, Claude-1/2, Vicuna (7B-v1.3, 13B-v1.3) (Chiang et al., 2023) and Mistral-7B (Jiang et al., 2023) using our EDA benchmarks. Unfortunately, we are unable to evaluate Llama-2 or its derivatives due to license restrictions. Llama-1 (Touvron et al., 2023a) is also excluded from the comparison as it lacks a conversation finetuned model, and as such it often does not adhere to our instructions. The corresponding results are presented in Table 1. The means and standard deviations for each dataset and LLM were calculated using 30 items from 5 separate runs.

In general, GPT-4 attained the best performance among the models evaluated. Around 1/3 of the entities were accurately predicted. The runner-ups are ChatGPT-3.5 and Claude-2. ChatGPT-3.5 is superior on the *Things* datasets, while Claude-2 shows better strength on the *Celebrities* datasets. Despite being much smaller, open-source models such as the Vicuna 13B and Mistral 7B model performed comparably well, not trailing too far behind the closed-source models. In particular, Vicuna 13B outperformed Claude-1 when it comes to *Things*. This suggests that open-sourced models have significant potential. Naturally, models with a higher success rate usually require fewer turns.

It is plausible that a small portion of entity deduction game sessions may exist on the web and were a part of the pretraining data. However, these sessions likely contribute only a minuscule fraction. Nevertheless, robust models such as GPT-4 exhibit consistently strong performance across numerous tested entities. We have also conducted a large-scale human baseline in a subset of 30 tested examples for each dataset. The discussion of these results can be found in Appendix B.

We present several real game runs in Table 2 on *Things*. Examples on *Celebrities* are in Appendix H. We observe that an effective game strategy typically featured by narrowing down the scope through initial questions that divide the remaining search space into two halves of similar capacity, similar to a *binary tree search* process. During the late game when it becomes challenging to find good binary-partitioning questions, the model start enumerates a smaller number of possibilities. Interestingly, the more advanced models like GPT-4 seem to possess this capability to a decent extent, while weaker models exhibit different failure modes in planning and reasoning abilities (Table 2): 1) (**Early Enumeration**) initiating enumeration too early without first posing high-level questions to narrow down the scope; 2) (**Redundancy**) asking questions similar to previous ones; 3) (**Inconsistency**) asking questions or making guesses that are inconsistent to the answers received in previous turns.

We also compare several games between GPT-4 and human players in Appendix G, highlighting that strong LLMs share similarities with humans in their ability to *backtrack* when they are on the wrong path. One advantage of LLM over humans is that LLM may have a superior level of knowledge (Appendix G, Table 8), empowering them to ask questions using *tacit knowledge* among LLMs. One weakness of the LLMs comparing to humans is that LLMs tend to enumerate or repeat to follow the same pattern of previous turns, demonstrating *self-reinforced* undesirable behaviors. Weaker models tend to be attracted to this absorbing state of repetition or artifacts, and accumulate errors until they ultimately fail (Appendix H). Stronger models occasionally display the ability to escape from this repetitive patterns and rectify early errors, which presumably be attributed to their stronger adherence to the high-level plan. We then address the following Research Questions (RQs):

RQ1: How does the LLM generate strategies? We conducted an analysis to probe the internal dialog states underneath the LLMs, to understand how these states evolve throughout the game, and how do they affect the strategy of the model. To this end, at each turn before the model asks a question, we prompt the LLM to predict top 5 candidates given

	Things				Celebrities			
	#Turns (↓)	Success (↑)	#Yes	Score (↑)	#Turns (↓)	Success (↑)	#Yes	Score (↑)
GPT-4	**17.8±0.2**	**0.31±0.03**	5.9±0.1	**0.26±0.02**	**17.3±0.1**	**0.50±0.02**	6.8±0.2	**0.40±0.02**
GPT-3.5	18.2±0.2	0.28±0.02	6.3±0.1	0.23±0.02	18.8±0.2	0.27±0.03	7.4±0.2	0.21±0.03
Claude-2	18.4±0.3	0.21±0.03	5.0±0.1	0.18±0.03	17.6±0.2	0.31±0.02	5.6±0.1	0.26±0.02
Claude-1	18.8±0.1	0.16±0.02	4.2±0.1	0.13±0.02	17.7±0.2	0.29±0.03	5.3±0.2	0.25±0.02
Vicuna 13B	18.4±0.1	0.18±0.02	5.0±0.2	0.15±0.02	18.7±0.2	0.22±0.03	6.1±0.1	0.18±0.02
Vicuna 7B	19.5±0.2	0.09±0.02	5.7±0.2	0.07±0.02	19.6±0.3	0.06±0.02	5.9±0.2	0.05±0.02
Mistral 7B	18.9±0.1	0.13±0.02	3.8±0.5	0.11±0.02	18.2±0.1	0.22±0.04	4.3±0.1	0.20±0.03
V-FT 7B (All)	19.2±0.1	0.13±0.01	6.1±0.1	0.10±0.01	19.3±0.1	0.16±0.02	7.6±0.3	0.13±0.02
V-FT 7B (Suc.)	18.0±0.1	0.23±0.01	5.1±0.2	0.20±0.01	19.0±0.2	0.15±0.02	6.3±0.2	0.13±0.02
V-FT 13B (All)	18.6±0.2	0.21±0.03	6.1±0.2	0.17±0.02	18.8±0.2	0.22±0.01	6.2±0.2	0.18±0.01
V-FT 13B (Suc.)	18.0±0.2	0.25±0.02	4.5±0.1	0.21±0.03	18.4±0.3	0.23±0.04	5.9±0.2	0.19±0.03
V-RLGP 7B	17.8±0.1	0.26±0.02	4.7±0.1	0.22±0.01	18.8±0.1	0.16±0.01	5.9±0.1	0.14±0.00
V-RLGP 13B	17.9±0.1	0.27±0.02	4.5±0.1	0.23±0.01	18.5±0.2	0.26±0.03	6.1±0.1	0.21±0.02

Table 1: Benchmark of LLMs on the EDA datasets (*Things*, *Celebrities*). **#Turns** denotes the average number of turns taken. **Success** denotes the percentage of the games that the G wins. **#Yes** denotes the average number of "yes" response received from J. **Score** indicates the reward score defined in Eq. (1). The mean and standard deviation were calculated based on 5 repetitions.

Successful game	Planning issue (Early Enumeration)
G: Is it a living entity? *J: No.* G: Is it man-made? *J: Yes.* G: Can it be held in a single hand? *J: Yes.* G: Is it electronic? *J: Yes.* G: Is it used for communication? *J: Maybe.* G: Can it store information? *J: Yes.* G: Is it a portable device? *J: Maybe.* G: Is it commonly found in homes? *J: Yes.* G: Is it related to entertainment? *J: No.* G: Is it used for work or productivity? *J: Yes.* G: Is it a type of computer or computing device? *J: Maybe.* G: Is it used for creating or editing documents? *J: Yes.* G: Is it a keyboard? *J: No.* G: Is it a mouse? *J: No.* G: Is it a printer? *J: **Bingo!***	G: Is the entity a living thing? *J: No.* G: Is the entity an object? *J: Yes.* G: Is the object a tool or a device? *J: Yes.* G: Is the tool or device used for **communication**? *J: No.* G: Is the tool or device used for **transportation**? *J: No.* G: Is the tool or device used for **construction**? *J: No.* G: Is the tool or device used for **entertainment**? *J: No.* G: Is the tool or device used for **cooking**? *J: No.* G: Is the tool or device used for **cleaning**? *J: No.* G: Is the tool or device used for **protection**? *J: No.* G: Is the tool or device used for **sports or athletics**? *J: No.* ...
Planning issue (Redundancy)	**Reasoning issue (Inconsistency)**
G: Is the entity a person? *J: No.* G: Is the entity a living thing? *J: No.* G: Is the entity a thing? *J: Yes.* G: Is the entity an object? *J: Yes.* G: Is the entity a material object? *J: Yes.* G: Is the entity a solid object? *J: Yes.* G: Is the entity a natural object? *J: No.* G: Is the entity an **artificial object**? *J: Yes.* G: Is the entity a **man-made object**? *J: Yes.* G: Is the entity an **inanimate object**? *J: Yes.* G: Is the entity a machine? *J: Yes.* ...	G: Is the entity a living being? *J: No.* G: Is the entity an object or a thing? *J: Yes.* G: Is the object typically found indoors? *J: Yes.* G: Is the object used primarily for a specific purpose? *J: Yes.* G: Is the object small and portable? *J: Maybe.* G: Is the object electronic? *J: Yes.* G: Is the object a device used for communication *J: No.* G: Is the object a device used for **entertainment** or media *J: No.* G: Is the object a device used for productivity or work *J: Yes.* G: Is the object a laptop? *J: No.* G: Is the object a **game console**? *J: No.* ...

Table 2: The gameplay with the hidden entity "printer". The successful and problematic gameplay scenarios are generated using GPT-4 and vicuna-7B model as G, respectively.

the current dialogue history. This probing process is separate from (thus will not affect) the gameplay. The probing prompt is provided in Appendix C. A GPT-4 game session of guessing "guitar" with dialog state is shown in Table 3. A session on *Celebrities* dataset of guessing "LeBron James" is shown in Table 12 in Appendix C.

From Table 3, it seems that the model seeks a strategy to ask questions to effectively partition the space among the most promising candidates in the initial 10 turns. After the question is answered, the ranked list of top guesses is adjusted, with certain top candidates being eliminated. For instance, during the third turn, the model asks about the item's feasibility of being held in one hand, leading to the elimination of the larger objects upon receiving an affirmative answer. During the fourth turn, the model asks if the item is electronic, which ultimately eliminates the top-ranked guess, "smartphone", after receiving a negative response. The remaining items are shifted one position forward to include a new item to the top guesses.

Interestingly, it is not necessary for the model to always ask questions in order to optimally partition the top-5 candidates. We observed that the strategy also depends on factors including the ease with

#	Current top 5 guesses before generating question					Generated question
1	computer	cellphone	book	car	house	G: Is it a living entity? J: No.
2	book	computer	car	building	smartphone	G: Is it man-made? J: Yes.
3	computer	car	building	book	smartphone	G: Can it be held in one hand? J: Yes.
4	smartphone	book	pen	key	mug	G: Is it electronic? J: No.
5	book	pen	key	mug	deck of cards	G: Is it used for writing or drawing? J: No.
6	key	deck of cards	coin	paper clip	dice	G: Is it made of metal? J: No.
7	paper book	wooden spoon	plastic toy	glass cup	porcelain plate	G: Is it used in cooking or eating? J: No.
8	book	playing card	toothbrush	comb	paper clip	G: Is it used for entertainment or leisure? J: Yes.
9	deck of cards	book	puzzle	board game	Rubik's cube	G: Is it associated with a specific sport? J: No.
10	book	deck of cards	dice	puzzle	board game	G: Is it used in board games? J: No.
11	deck of cards	book	paintbrush	puzzle	dice	G: **Is it associated with music?** J: Yes.
12	Harmonica	Guitar pick	Ocarina	Castanets	Recorder	G: Is it a musical instrument? J: Yes.
13	Harmonica	Recorder	Kazoo	Maracas	Tambourine	G: **Is it a stringed instrument?** J: Yes.
14	Ukulele	Violin	Guitar	Mandolin	Harp	G: Is it commonly used in a symphony orchestra? J: No.
15	Guitar	Ukulele	Banjo	Mandolin	Lute	G: Is it a guitar? J: Bingo!

Table 3: Probing the top 5 ranked guesses made by GPT-4 during a game of guessing "Guitar". Some guesses were truncated for readability. **Highlighted turns** are the key turns that eliminates all top 5 guesses.

which a question can be asked to partition the space, and the level of uncertainty the model has about the current top predictions. We provide details on the measurement of uncertainty in Appendix D. In situations where the model is uncertain, it may occasionally *backtrack* and *reexamine* categories that were previously missed. For instance, during the 11th turn, question was asked that ruled out all of the top candidates. This could be due to the fact that the top items are similar in nature, and the model realize there is a significant proportion of other classes that have been overlooked. The successful questioning in these cases led to the recovery of these overlooked classes.

	Things (↑)	Celebs (↑)
GPT-4 → Vicuna 7B	0.10±0.02	0.14±0.04
GPT-4 → GPT-4	**0.26±0.02**	**0.40±0.02**
Vicuna 7B → Vicuna 7B	0.07±0.02	0.05±0.02
Vicuna 7B → GPT-4	0.08±0.01	0.06±0.01

Table 4: Ablation on planning and reasoning ability. Numbers are the game scores with 5 repetitions. X → Y indicates using the conversation trajectory from X and use the Y for the last turn.

To summarize, a strong G exhibits the following behavior: 1) **prioritizing high-level questions** before addressing specific details and enumerations, 2) being aware of the current state and asking questions to effectively **bi-partition the search space**, and 3) being able to **occasionally backtracking** to consider previously overlooked options.

Additionally, we noticed a high level of consistency in GPT-4's gameplay strategy across multiple repetitions (see example in Appendix L), despite some fluctuation in the order of the questions asked. This suggests that GPT-4 may rely on its own implicit *taxonomy representation* of entities to make decisions, which remains consistent throughout.

RQ2: Which one is more important in this task, planning or reasoning? Planning and reasoning abilities affects different stages of game play. Early questions require careful planning to efficiently partition the space, while late game requires more deductive reasoning skills to make an educated guess. We consequently assume that the last turn would only require reasoning ability as no strategic move is needed. With this assumption, we designed the following experiment to investigate the model's planning and reasoning ability in a finer granularity. Given a stronger model GPT-4 and a weaker model Vicuna 7B, and their respective game play trajectories, we only replay the last turn in each trajectory by swapping the guesser model G.

The results are presented in Table 4. Comparing the GPT-4 → Vicuna 7B with Vicuna 7B, we observe that stronger planning ability from GPT-4 results in significant improvement. Regarding different reasoners, Vicuna 7B → GPT-4 does not show much improvement over Vicuna 7B. This indicates that planning deficiency could result in an unproductive trajectory, poses significant challenges for reasoning during the final step. Moreover, GPT-4 → Vicuna 7B shows regression over GPT-4, emphasizing the importance of reasoning in addition to a strong planning capability. Therefore, it is crucial for both planning and reasoning abilities to be strong and work in synergy to achieve optimal performance. See the full table including the results using human trajectory in the Appendix E.

4 Enhancing the Open-source Models

We further investigate the potential of open-source models. We used Behavior Cloning (BC) to distill capabilities exhibited by the stronger models into the smaller Vicuna models. We first collected game demonstrations from the GPT-3.5 model over all 600 training examples from *Things* and *Celebrities*. We then fine-tuned the Vicuna models using these demonstrations. The experimental settings are provided in the Appendix I.

RQ3: Can smaller open-source models benefit from imitating larger close-source models? We first fine-tuned the Vicuna 7B model using a total of 600 training instances from both datasets. The results are presented in Table 1 as the **V-FT 7B (All)** model. This yielded a more than 70% improvement in both datasets, indicating that weaker models can follow the stronger models' demonstration to use a more effective strategy to steer the conversation on unseen entities or names. We have also observed that BC training can effectively mitigate undesirable behavior such as repetition.

RQ4: Should models learn from successful demonstration exclusively? To answer this, we selected all the game runs of the entities or names where the teacher model finally won the game from both datasets. This subset was then used to train the Vicuna 7B model, similar to the *Rejection Sampling* (Touvron et al., 2023b) . The results, referred to as **V-FT 7B (Suc.)** in Table 1, show that imitating from successful experiences outperforms imitating from all experiences (**V-FT 7B (All)**). It gets close to the teacher model (GPT-3.5) on *Things*. However, when it comes to the *Celebrities* dataset, the improvement is more limited.

RQ5: Does the model size matter? We further compared the performance improvement achieved through finetuning for both Vicuna 7B and 13B models. The model finetuned on Vicuna 13B, **V-FT 13B (Suc.)**, also achieved much higher score than Vicuna 13B. However, the improvement was less substantial than the improvement on **V-FT 7B (Suc.)**. Again, the improvement over the *Celebrities* dataset is marginal. We hypothesize that guessing the celebrity requires a more case-specific strategies compared to guessing things, thus likely to be more challenging to be superficially imitated.

5 Reinforcement Learning via Gameplay

We conducted further experiments to investigate whether the performance could be further enhanced through learning solely from the model's own experience. We employed Proximal Policy Optimization (PPO) (Schulman et al., 2017) to train the model by playing with with the judge, J, a technique we refer to as Reinforcement Learning from Game-play RLGP. We made modifications to the trlX repository (Castricato et al., 2023) to enable multiturn setup to facilitate RLGP training. During training, we assigned the reward defined in Eq. (1), to the final turn of each rollout. Additionally, we assigned an intermediate reward to turns that received a "Yes" response from J. We trained the V-FT 7B (Suc.) and V-FT 13B (Suc.) models in Table 1 on 300 *Things* training dateset. Further experimental details can be found in Appendix K. Table 1 presents the performance of the RL-trained models, denoted as **V-RLGP**.

V-RLGP models exhibit improvement compared to the vanilla Vicuna when tested on the in-domain dataset *Things*. **V-RLGP 13B** seems to unlock the potential of the Vicuna model, outperforming the **V-FT 13B** model, matching the performance of the runner-up GPT-3.5. **V-RLGP 7B** also show substantial improvement over **V-FT 7B** model. On the out-domain dataset *Celebrities*, interestingly, **V-RLGP** models achieves some improvement, showing its generalization potential.

6 Breakdown Analysis: Do the models agree on successful predictions?

We present a comprehensive breakdown analysis of each model's performance on a subset of 30 evaluated items for each dataset in Figure 2. Within the evaluated items of *Things*, four items consistently defy successful predictions by all models. On the other hand, *Celebrities* exhibits a more scattered pattern, with each celebrity being correctly predicted by at least one model.

There are correlations between the entities or names that each model can correctly identify, but different models exhibit their own strengths on different subsets. For instance, the **RLGP 13B** model can accurately predict "Sunglasses", whereas the stronger models like GPT-4 consistently fails. We also provide some case studies including why GPT-4 consistently fails on "Yoga mat" in Appendix L.

We observed that RLGP models tend to

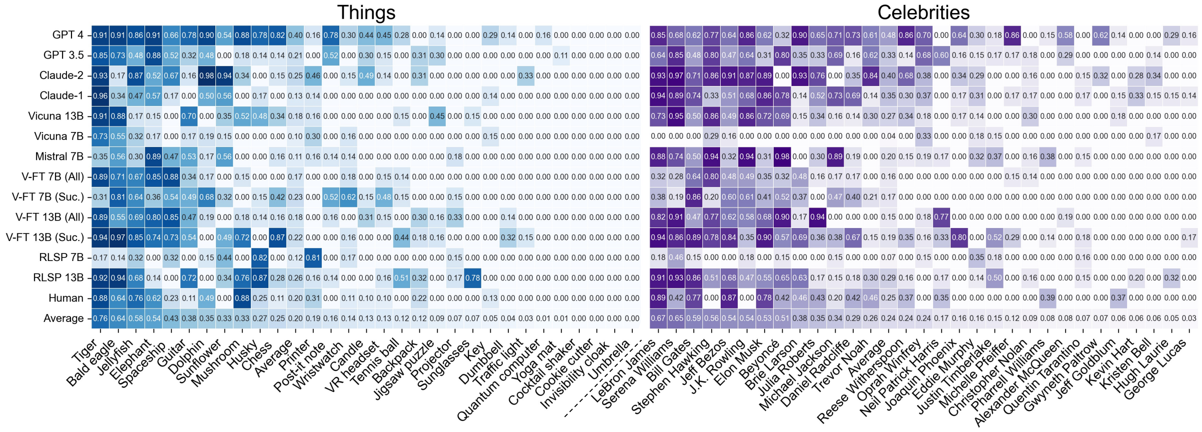

Figure 2: A breakdown of the score of each model on the evaluated items, with the x-axis representing the order of difficulty ranging from easy to difficult. Scores are averaged over 5 repetitions.

strengthen the performance on items that vanilla models occasionally succeed in, thereby improving their success rate on these specific items. However, RLGP models do not effectively facilitate learning about new items. Conversely, BC fine-tuning excel in achieving success on new items. Interestingly, the BC fine-tuned models displays different strengths compared to both the initial checkpoint and the expert whose demonstration it mimics. For example, the **V-FT 13B (Suc.)** model achieves high accuracy in identifying Post-it Note and VR headset, whereas neither the Vicuna 13B nor the GPT-3.5 performs as well in this regard.

7 Related Work

Complex Reasoning Benchmarks Extensive benchmarks have been evaluating the complex reasoning abilities of LLMs (Huang and Chang, 2022). These include HELM (Liang et al., 2022) BIG-bench (Srivastava et al., 2022), SuperGLUE (Sarlin et al., 2020), LAMA (Petroni et al., 2019), and CoT-Hub (Fu et al., 2023), which have enabled researchers to assess LLMs across a spectrum of tasks involving arithmetic and math (GSM8K) (Cobbe et al., 2021), commonsense (ARC) (Clark et al., 2018), knowledge (MMLU) (Hendrycks et al., 2020), and coding (HumanEval) (Chen et al., 2021) have been developed to gauge their real-world abilities. We aim to introduce a novel benchmark focus on evaluating multi-turn state tracking and strategical planning capability to complement existing reasoning benchmarks.

Evaluation of Planning Evaluation of the planning abilities of LLMs is relatively rare. Valmeekam et al. (2022) proposed an assessment framework to gauge the planning capabilities of LLMs to generate valid actions to achieve a special goal and provide a rationale for the plan. Their evaluation on 8 planning tasks reveals LLMs, like GPT-3 (Brown et al., 2020), seem to display a dismal performance. Valmeekam et al. (2023) further evaluates on GPT-4 and suggests the autonomous learning capacity of LLMs to formulate plans is limited and dependent on properly designed heuristics. Xie et al. (2023) similarly indicate that LLMs may encounter difficulties in generating planning goals related to numerical or spatial reasoning, while being responsive to the specific prompts used. Unlike these studies, our task solely relies on textual representations of goals.

Multiturn benchmarks MT-Bench (Zheng et al., 2023) assesses the multi-turn conversation and instruction-following ability of LLMs by annotating their responses to questions involving writing, role-play, extraction, reasoning, math, coding and knowledge. Bang et al. (2023) evaluates LLMs on 23 tasks related to logical reasoning, non-textual reasoning, and commonsense reasoning. The study reveals that incorporating interactive features in LLMs can enhance their performance by employing multi-turn prompt engineering. LMRL Gym (Abdulhai et al., 2023), a concurrent work, investigated Multi-Turn Reinforcement Learning with Language Models across multiple tasks. In contrast, our work focuses on the depth by presenting a thorough analysis of the entity deduction task.

Entity-deduction game Testing the model's ability to deduct an ambiguous entity or asking clarification questions (Aliannejadi et al., 2019; Cho et al., 2019) has been utilized as a testbed on dialogue systems and visual reasoning tasks. InfoBot (Dhingra et al., 2016) uses reinforcement learning to learn an optimal dialog policy for identifying movie entries from a movie database. ESP (Von Ahn and Dabbish, 2004) and Peekaboom (Von Ahn et al., 2006) demonstrated that deduction games can effectively gather labeled data. GuessWhat?! (De Vries et al., 2017) and ReferIt (Kazemzadeh et al., 2014), assess the visual reasoning capabilities of tested models by ask them to guess the referred object in the image based on the conversation. Our work instead aims to gauge on the model's ability on generating the conversation.

8 Conclusion

Motivated by a need to develop agents capable of effectively addressing ambiguous user intents, we introduce a testbed for evaluating LLM's strategic planning and deductive reasoning abilities in asking entity-deducing questions. Our findings indicate that SOTA LLMs are able to maintain an intrinsic taxonomic representation of knowledge entities to a certain extent. We further show that this capability can be enhanced through Behavior Cloning or Reinforcement Learning, revealing great potential for further advancements. In future research, we intend to investigate whether the implementation of CoT prompting can further enhance the model's performance in related tasks.

9 Limitations

The Entity-Deduction Arena presents a narrow aspect of LLM multiturn planning capabilities, focusing on iterative refinement over selective taxonomies. As such, the results of this work presents a peak into LLM planning behavior, but doesn't necessarily apply to all LLM planning problems.

Despite our best efforts to ensure the diversity of the datasets, they could still be subject to bias such as origin language, popularity and time relevance due to their web-crawled nature.

Finally, while Behavior Cloning and RLGP are shown to be useful to improve the performance of Vicuna models, future work could measure their effectiveness across a variety of open source model families.

References

Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. 2023. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models.

Akinator. 2007. Akinator. Accessed on September 7, 2023.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Louis Castricato, Alex Havrilla, Shahbuland Matiana, Duy V. Phung, Aman Tiwari, Jonathan Tow, and Maksym Zhuravinsky. 2023. trlx: A Scalable Framework for Rlhf.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Woon Sang Cho, Yizhe Zhang, Sudha Rao, Asli Celikyilmaz, Chenyan Xiong, Jianfeng Gao, Mengdi Wang, and Bill Dolan. 2019. Contrastive multi-document question generation. *arXiv preprint arXiv:1911.03047*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2016. Towards end-to-end reinforcement learning of dialogue agents for information access. *arXiv preprint arXiv:1609.00777*.

Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. 2023. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv preprint arXiv:2305.17306*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Langchain-AI. 2023. langchain Github Repository. GitHub repository.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? In *EMNLP*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.

Significant Gravitas. 2023. auto-gpt: An Autonomous Gpt-4 Experiment. GitHub repository.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models–a critical investigation. *arXiv preprint arXiv:2305.15771*.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*.

Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326.

Luis Von Ahn, Ruoran Liu, and Manuel Blum. 2006. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 55–64.

Yaqi Xie, Chen Yu, Tongyao Zhu, Jinbin Bai, Ze Gong, and Harold Soh. 2023. Translating natural language to planning goals with large-language models. *arXiv preprint arXiv:2302.05128*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Appendix

A Dataset composition

We provide the dataset composition on *Things* and *Celebrities* in Table 3.

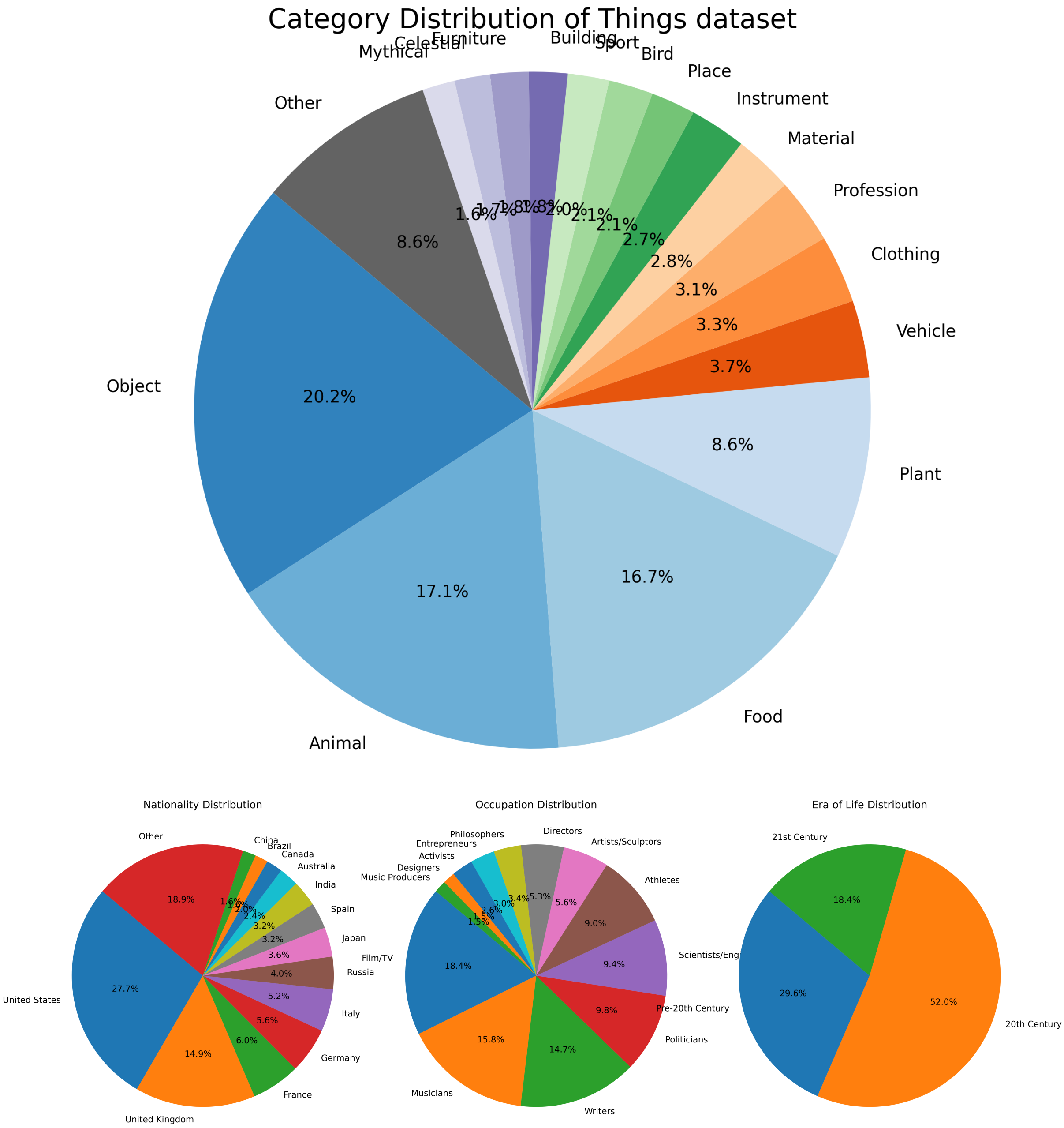

Figure 3: Composition of EDA *Things* and *Celebrities* datasets.

B Prompt used for *Celebrities* and rationale for using "Dunno"

For *Celebrities*, the J use the following prompt to generate a response.

Based on on your knowledge about the celebrity: {entity} respond to the following question or guess. Limit your respond to only 'Yes.', 'No.' or 'Dunno.', with no explanation or other words. Never say the name {entity} in your response. Do not say 'Dunno.' if it can be answered by 'Yes.' or 'No.' If the question is to solicit the answer, respond 'No.'.\n Question/Guess: {question}(Yes/No/Dunno)

The G use the following prompt to generate a question or guess.

Your task is to ask a series of questions to deduce the celebrity that I'm thinking of with as few queries as possible. Only ask questions that can be answered by "Yes", "No", or "Dunno". Do not ask for hint. Make your question brief with no linebreaker. Now start asking a question. \n {dialog history}

Rationale for using "Dunno" The Things dataset contains entities that has many different forms, such as if asking "is this made by metal?" for the entity "Chopstick". The answer to this question could be "Maybe" because sometimes they are made by metal and sometimes are not. On the other hand, in the Celebrities dataset, most questions can typically be clearly answered with a yes or no. These questions often refer to gender, birth year, or other specific information about a person. However, there is a possibility that the model might not know the answer to certain questions, such as "Does this celebrity smoke?". In these cases, the answer could be "Dunno" as it is considered very personal information. As a general observation, we have found that less than 1% of the responses in the Celebrities dataset are "Dunno", while around 20% of the responses in the Things dataset are "Maybe". Note that the percentage of "Dunno" or "Maybe" responses can vary depending on the specific G being used.

Human baseline Collecting static human annotation for this study is a challenging task due to the interactive nature of this research. In order to establish a baseline of human performance, we conducted a human-in-the-loop study. We subsampled *Things* and *Celebrities* into 2 smaller subsets containing 30 sample datapoints each, *Things-30* and *Celebrities-30*. We set up a game server and recruited 108 non-paid internal volunteers to interact with the J, and collected a total of 145 and 71 human game play sessions for *Things-30* and *Celebrities-30*, respectively. Human guessers were given the same instructions as the LLM guessers and were provided with a tutorial. Optionally, for training, they could also view the question GPT-3.5 would have asked at the previous step. Statistics, experimental details and UI are provided in Appendix F.

To ensure the quality of the data, we manually inspected human game plays to filter out sessions that do not complete the entity deduction task, contain irrelavant chit-chat, or contain nonsensical repeated questions, leaving only sensible game plays. In the end 201 sessions (93.1% of the total game play) were selected as valid game play.

C Additional results and prompt used for Dialog State Probing

We used the following prompt to probe the model for its top guesses during each turn.

{dialog history}
Based on the information provided, here are the top 5 most likely concrete entities I think you are thinking of:

We provide additional top K probing results from GPT-4 on *Celebrities* in Table 12. Similarly, the model seems to be able to leverage its own hierarchical representation of the celebrities' space to ask questions that steer to narrow down the search effectively.

	Things-30				Celebrities-30			
	#Turns (↓)	Success (↑)	#Yes	Score (↑)	#Turns (↓)	Success (↑)	#Yes	Score (↑)
GPT-4	**16.9±0.2**	**0.49±0.06**	6.0±0.2	**0.40±0.05**	16.5±0.5	**0.59±0.04**	7.3±0.1	**0.48±0.03**
GPT-3.5	18.4±0.3	0.25±0.04	7.1±0.4	0.21±0.04	17.9±0.3	0.41±0.05	7.6±0.3	0.33±0.04
Claude-2	17.6±0.3	0.29±0.05	4.5±0.3	0.25±0.04	**15.9±0.4**	0.45±0.06	5.3±0.1	0.40±0.05
Claude-1	18.7±0.1	0.15±0.02	4.3±0.2	0.13±0.02	16.7±0.4	0.41±0.05	4.6±0.2	0.35±0.04
Vicuna 13B	18.7±0.2	0.20±0.03	5.2±0.3	0.17±0.02	17.7±0.4	0.36±0.08	6.8±0.3	0.27±0.06
Vicuna 7B	19.1±0.4	0.11±0.06	5.7±0.6	0.10±0.05	19.7±0.3	0.05±0.04	6.2±0.7	0.04±0.03
V-FT 7B (All)	18.4±0.2	0.20±0.02	6.8±0.2	0.17±0.02	19.0±0.2	0.21±0.04	9.1±0.3	0.16±0.03
V-FT 7B (Things)	18.5±0.4	0.22±0.06	6.6±0.2	0.18±0.05	19.1±1.5	0.19±0.20	10.3±3.6	0.15±0.17
V-FT 7B (Celebs)	19.7±0.3	0.03±0.02	1.6±0.1	0.03±0.02	19.1±0.2	0.20 ±0.07	7.5±0.6	0.16±0.05
V-FT 7B (Suc.)	18.5±0.5	0.28±0.10	6.8±0.5	0.23±0.08	18.6±0.5	0.21±0.06	7.4±1.4	0.17±0.04
V-FT 13B (Suc.)	18.0±0.5	0.29±0.08	6.9±0.2	0.24±0.07	18.6±0.6	0.22±0.09	7.8±0.5	0.18±0.07
V-RLGP 7B	19.3±0.2	0.15±0.03	3.6±0.1	0.12±0.02	19.5±0.3	0.09±0.05	5.8±1.1	0.07±0.04
V-RLGP 13B	17.8±0.2	0.31±0.03	4.0±0.2	0.26±0.02	17.5±0.5	0.35±0.04	6.8±0.2	0.29±0.04
Human	18.5±0.5	0.24±0.04	5.2±0.2	0.20±0.04	18.1±0.2	0.31±0.03	7.0±0.3	0.25±0.03

Table 5: Benchmark of LLMs on the EDA datasets (*Things-30*, *Celebrities-30*) including human baseline. **#Turns** denotes the average number of turns taken. **Success** denotes the percentage of the games that the `G` wins. **#Yes** denotes the average number of "yes" response received from `J`. **Score** indicates the reward score defined in Eq. (1). The mean and standard deviation were calculated based on 5 repetitions.

D Measurement of uncertainty

We asked the Guesser model, "On a scale of 1-5, how confident are you in your top guesses based on the current conversation history?" We observed a correlation between the model's predicted uncertainty and significant shifts in its top guesses. However, this pattern was mainly observed in more powerful models such as GPT-4. In the case of weaker models like Vicuna, the top guesses or confidence scores generated by the model were not very sensible.

E Full ablation results on RQ2, planning vs reasoning

	Things-30 (↑)	Celebs-30 (↑)
GPT-4 → Vicuna 7B	0.12±0.03	0.19±0.02
GPT-4 → GPT-4	**0.40±0.05**	**0.48±0.03**
Vicuna 7B → Vicuna 7B	0.10±0.04	0.04±0.03
Vicuna 7B → GPT-4	0.11±0.03	0.06±0.03
Human → Vicuna 7B	0.11±0.04	0.03±0.01
Human → GPT-4	0.15±0.01	0.18±0.06
Human → Human	0.20±0.04	0.25±0.03

Table 6: Ablation on planning and reasoning ability. Note that non-human game plays are also evaluated on the subsampled datasets *Things-30* and *Celebrities-30* for a fair comparison. Numbers are the game scores with 5 repetitions. X → Y indicates using the conversation trajectory from X and swap the Y for the last turn. For example, GPT-4 → Vicuna 7B uses GPT-4 to play all except the last turn, swapping in Vicuna 7B in the last turn.

The full results comparing planning and reasoning are presented in Table 6. Interestingly, when GPT-4 is employed for the last step reasoning in human trajectories, there is a decrease in performance. One possible explanation for this is that while GPT-4 may excel in comprehending model-generated trajectories, it may struggle to fully interpret the planning trajectory of humans based on the observed turns. This suggests that models may be more compatible with model-generated trajectories rather than human trajectories.

F Experimental details for human baseline

In total, we recruited 108 human volunteers to participate in the study. To ensure the quality of the data, we manually inspected human game plays to filter out low quality data. We also prioritized games that received fewer qualified plays to present to the human players. After a 30-day trial period, we collected a total of 140 and 68 human game play sessions for *Things* and *Celebrities*, respectively. Each entity or celebrity is covered at least by two game plays.

We show the UI for the in game demo server in Figure 4. The UI contains 3 sections – game play, tutorial and leaderboard. In the leaderboard, we compute the Wilson confidence interval for each player, and include the performance of LLMs as benchmarks for the human players to refer to.

For Human game plays, entities with fewer game plays are oversampled so that all entities have the same amount of game play. Game plays are then randomized into separate runs to calculate means and standard deviations. Note that this might not be exactly comparable to the model's numbers.

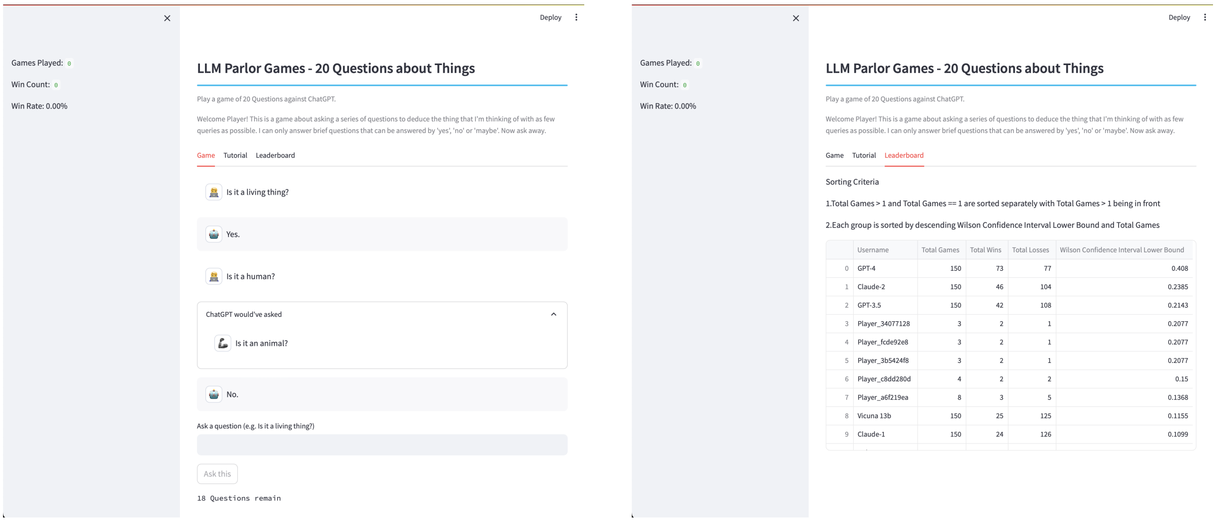

Figure 4: Game play UI interface for collecting human baseline. On the left, human players are given prompt instructions equivalent to LLM guessers. An optional retrospection UI can be toggled to display what ChatGPT would've chosen to ask in the last turn. On the right, a leaderboard with Human and LLM player performance is shown.

G Comparison of GPT-4 vs Human player

We provide additional examples in Table 7 (*Things*) and Table 8 (*Celebrities*) for GPT-4's generation and compare them with the human players, highlighting that strong LLMs share similarities with humans in their ability to *backtrack* when they are on the wrong path. One advantage of LLM over humans is that LLM may have a superior level of knowledge (Table 8), empowering them to ask questions using *tacit knowledge* among LLMs. One weakness of the LLMs comparing to humans is that LLMs tend to enumerate or repeat to follow the same pattern of previous turns, demonstrating *self-reinforced* undesirable behaviors. Weaker models tend to be attracted to this absorbing state of repetition or artifacts, and accumulate errors until they ultimately fail (Appendix H). Stronger models occasionally display the ability to escape from this repetitive patterns and rectify early errors, which presumably be attributed to their stronger adherence to the high-level plan. We then present a detailed analysis on each dataset:

G.1 *Things* dataset

In the first example "Guitar", we observed that both GPT-4 and the human player efficiently converged on a music-related object within approximately 10 turns. However, the human player deviated from the correct path by thinking about an electronic device, whereas GPT-4 promptly identified the item as a musical instrument. Notwithstanding this, the human player's capacity for back-tracking their steps was evident when they recognized the erroneous direction and ultimately win the game. In this game, the human player retraced the error and asking a reconfirming question "Is it some kind of music player?".

We then ask the question if the GPT-4 model also has this back-tracking ability. The answer is yes. In Table 7 (Guitar, Human + GPT-4), we ask GPT-4 to complete the game based on partial gameplay provided by human players until they recognize their mistake (up to "G: Can it display images? J: No."). GPT-4 promptly comprehends the game scenario and identifies that it has taken the incorrect route. Subsequently, it formulates the appropriate question that encompasses the overlooked category, thereby salvaging the game. We have also observed this similar capability in several other game sessions from GPT-4.

In the second game "Bald eagle", human player has a luck guess after identifying the entity is a bird. GPT-4, on the contrary, still roll out 3 more turns to gather more information before making an informative guess.

In the third example "Jigsaw Puzzle", GPT-4 interleaves questions regarding size, purpose and location and combine multiple aspects of information to make a guess. Prior to making the final guess, GPT-4 demonstrates caution by reconfirming if the object is a physical toy, after receiving a "No" with "Is it a board game?". On the contrary, the human player seems clueless and may ask repetitive or irrelevant questions.

G.2 *Celebrities* dataset

We present a comparative analysis between GPT-4 and human players on *Celebrities* in Table 8. Our findings reveal that GPT-4 demonstrates a propensity for asking questions that necessitate domain-specific background knowledge (as highlighted in Table 8). Such knowledge could be *tacit knowledge* between LLMs, which human players may lack. With its extensive information repository, GPT-4's inquiries can delve into highly specific topics, sometimes requiring professional expertise that human judges may not possess. This characteristic confers an additional strength and advantage to GPT-4 during gameplay. Conversely, human players tend to rely more on direct guesses, which may yield fortunate outcomes in certain scenarios, but limits their overall effectiveness when compared to GPT-4.

H Game runs and failure modes on *Celebrities* and *Things*

We provide additional gameplay sessions on *Celebrities*, as shown in Table 9. Similar findings can be observed as they are in the experiment section. We observed that stronger models display proficient gameplay by accurately identifying the relevant question that separates the current space, while weaker models encounter difficulties in planning and reasoning.

We also presents several other failure modes observed with the Vicuna 7b model in Table 10. These failure modes illustrate the repetitive behavior and generation of artifacts. When a less robust model is drawn towards an absorbing state of repetitive pattern, it becomes increasingly challenging for them to escape from it.

I Experimental settings for Bahavior Cloning

We used 4x A100 Nvidia GPU for training. The learning rate was configured to be 2e-5, while the batch size was set to 32. We utilized the BF16 method for training and consistently employed deepspeed stage 2 across all our experiments. As part of the training process, we masked tokens from the turns of the J within the loss function. In all of our experiments, we continued training the model until the validation loss ceased to decrease any further.

J Generalization ability of Vicuna 7B finetuned on *Things* to *Celebrities*

A comparison of the gameplay between Vicuna 7B and the **V-FT 7B (Things)** is provided in Table 11. We observe that Vicuna 7B model may be more inclined to ask irrelevant questions, while the V-FT 7B (things) ask questions directly to the point. In terms of game strategy, the V-FT 7B (things) is more generative effective, indicating the planning ability obtained from imitation might be able to generalize to out of domain scenario.

K Experimental settings for Reinforcement Learning from Game-Play (RLGP)

We adhered to the default settings of the TRLX library for our experiments, but made modifications to certain hyperparameters. Specifically, we set the rollout generation temperature to 0.8 and disabled the topK or topP sampling. The value function coefficient for the value hydra head was set to 0.05. For the discounted reward, we set γ to 1 and λ to 0.97. The RL models were trained using four Nvidia A100 cards, with a total of 32 rollouts for each training entity. We trained the Vicuna 7B and 13B models on a dataset of 209 entities from the *Things* domain, which at least one of the off-the-shelf LLMs had won at least once. This is to ensure that model learn from experiences where it has some chance to receive some final rewards.

The intermediate reward was determined as follows. We utilized a linear decaying function to gradually decrease the intermediate reward from 2 to 0, based on the following equation:

$$R_{\text{intermediate}} = \max(0.2 - 0.025 \cdot \textbf{\#Turns}, 0) \quad (2)$$

The rationale behind this choice is that achieving more "yes" responses in the earlier turns is considered a positive signal. We observed that this approach improved the performance of the RLGP model. If the initial intermediate reward is too high (*e.g.,* 0.5), the model tends to generate trivial questions like "is it an entity" more frequently. Our choice is based the empirical results obtained from the development set. All RLGP models were trained for 600 iterations, ensuring that the model does not deviate significantly from the initial Vicuna model.

L GPT-4 failure cases

We notice that the strongest model GPT-4 consistently fail on the entity "Yoga Mat", while some other models can adeptly deduct it. We investigated this issue and it turns out that the GPT-4 model will always ask a misleading question "Is it made of plastic" in the 5 tested runs, leading to self-misinterpretation. See examples in Table 13. GPT-4 seems to possess a stereotype of plastic items and excludes "Yoga mat" after knowing this item is made of plastic. The effect of this question is more evident in Table 14. After confirming that the entity is made of plastic, the model proceeds to focus on plastic only items in its top guesses for all following turns. This observation suggests that even the most robust model, which may demonstrate backtracking when encountering a wrong path, can still become trapped in erroneous assumptions.

Item: Guitar		
GPT-4	**Human player**	**Human + GPT-4**
G: Is it a living entity? *J: No.* G: Is it man-made? *J: Yes.* G: Can it be held in one hand? *J: Yes.* G: Is it electronic? *J: No.* G: Is it used for writing or drawing? *J: No.* G: Is it made of metal? *J: No.* G: Is it used in cooking or eating? *J: No.* G: Is it used for entertainment or leisure? *J: Yes.* G: Is it associated with a specific sport? *J: No.* G: Is it used in board games? *J: No.* G: Is it associated with music? *J: Yes.* G: Is it a musical instrument? *J: Yes.* G: Is it a stringed instrument? *J: Yes.* G: Is it commonly used in a symphony orchestra? *J: No.* G: Is it a guitar? *J: Bingo!*	G: Is it a living thing? *J: No.* G: Is it a man-made object? *J: Yes.* G: Is it commonly used indoors? *J: Yes.* G: Is it larger than a toaster? *J: Yes.* G: Is it used for recreational purposes? *J: Yes.* G: Does it require electricity? *J: Maybe.* G: Does it have a screen? *J: No.* G: Does it contain metal? *J: Yes.* G: Does it contain glass? *J: No.* G: Does it make sound? *J: Yes.* G: Is it used to play music? *J: Yes.* G: Is it a speaker? *J: No.* G: Is it a radio? *J: No.* G: Does it have a wireless connection? *J: No.* G: Can it display images? *J: No.* G: **Is it some kind of music player?** *J: No.* G: Is it a musical instrument? *J: Yes.* G: Is it a guitar? *J: Bingo!*	G: Is it a living thing? *J: No.* G: Is it a man-made object? *J: Yes.* G: Is it commonly used indoors? *J: Yes.* G: Is it larger than a toaster? *J: Yes.* G: Is it used for recreational purposes? *J: Yes.* G: Does it require electricity? *J: Maybe.* G: Does it have a screen? *J: No.* G: Does it contain metal? *J: Yes.* G: Does it contain glass? *J: No.* G: Does it make sound? *J: Yes.* G: Is it used to play music? *J: Yes.* G: Is it a speaker? *J: No.* G: Is it a radio? *J: No.* G: Does it have a wireless connection? *J: No.* G: Can it display images? *J: No.* GPT: Is it a musical instrument? *J: Yes.* GPT: Is it a string instrument? *J: Yes.* GPT: Is it a guitar? *J: Bingo!*

Item: Bald eagle	
GPT-4	**Human player**
G: Is it a living thing? *J: Yes.* G: Is it an animal? *J: Yes.* G: Is it a domestic animal? *J: No.* G: Is it a mammal? *J: No.* G: Is it a reptile? *J: No.* G: Is it a bird? *J: Yes.* G: Is it a bird of prey? *J: Yes.* G: Does it have a predominantly white color? *J: Yes.* G: Is it found in the Polar regions? *J: No.* G: Is it a bald eagle? *J: Bingo!*	G: Is it alive? *J: Yes.* G: Does it live on land? *J: Yes.* G: Is it a plant? *J: No.* G: Is it human? *J: No.* G: Does it live in the Americas? *J: Yes.* G: Does it live in North America? *J: Yes.* G: Is it a predator? *J: Yes.* G: Is it a bear? *J: No.* G: Is it a bird? *J: Yes.* G: Is it a bald eagle? *J: Bingo!*

Item: Jigsaw Puzzle	
GPT-4	**Human player**
G: Is it a living entity? *J: No.* G: Is it man-made? *J: Yes.* G: Is it used for communication? *J: No.* G: Is it larger than a microwave? *J: Maybe.* G: Is it used for transportation? *J: No.* G: Is it made of metal? *J: No.* G: Is it used in homes? *J: Yes.* G: Is it used for sitting? *J: No.* G: Is it used for storage? *J: No.* G: Is it a decorative item? *J: Maybe.* G: Is it used in the kitchen? *J: No.* G: Is it used in the living room? *J: Maybe.* G: Is it used for entertainment purposes? *J: Yes.* G: Is it a musical instrument? *J: No.* G: Is it a television? *J: No.* G: Is it a game or a toy? *J: Yes.* G: Is it a board game? *J: No.* G: Is it a video game console? *J: No.* G: Is it a physical toy? *J: Yes. You must guess now, what's it?* G: Is it a puzzle? Is it right? *J: Yes.*	G: What is the meaning of life? *J: No.* G: What is a good thing? *J: Maybe.* G: Is it human? *J: No.* G: Is it an object? *J: Yes.* G: Is it a tool? *J: Yes.* G: Is it a hammer? *J: No.* G: Is it a kitchen tool? *J: No.* G: What is it? *J: Maybe.* G: Does it cut? *J: No.* G: Is it for food? *J: No.* G: Is it for cars? *J: No.* G: Is it big? *J: Maybe.* G: Can you eat it? *J: No.* G: Is it red? *J: No.* G: Is it metal? *J: No.* G: Is it used for communication? *J: No.* G: Is it used for storing? *J: No.* G: Is it used for helping humans? *J: Maybe.* G: Is it used for taking pictures? *J: No.* G: Alarm clock *J: No.*

Table 7: The gameplay scenario generated using GPT-4 model vs the gameplay scenario by human players, on *Things*.

GPT-4	Human player
Celebrity: Reese Witherspoon	
G: Is the celebrity you're thinking of alive? J: *Yes.* G: Is this celebrity a man? J: *No.* G: Is this celebrity an actress? J: *Yes.* G: Is she American? J: *Yes.* G: Has she won an Oscar? J: *Yes.* G: Is she over 50 years old? J: *No.* G: Has she starred in a superhero movie? J: *No.* G: Is she blonde? J: *Yes.* G: **Has she appeared in a film directed by Martin Scorsese?** J: *No.* G: Has she won her Oscar within the last 10 years? J: *No.* G: Is she known for romantic comedy films? J: *Yes.* G: Is it Reese Witherspoon? J: *Bingo!*	G: Male? J: *No.* G: White? J: *Yes.* G: American J: *Yes.* G: American J: *Yes.* G: Actor J: *Yes.* G: Alive J: *Yes.* G: Action movies J: *No.* G: Romantic comedies J: *Yes.* G: Sandra bullock J: *No.* G: Over 50 years old J: *No.* G: Emma stone J: *Dunno.* G: Is the person female? J: *Yes.* G: Meg Ryan J: *No.* G: Amy Schumer J: *No.* G: Kate Hudson J: *No.* G: Julia Roberts J: *No.* G: Cameron Diaz J: *No.* G: Reese Witherspoon J: *Bingo!*
Celebrity: Trevor Noah	
GPT-4	**Human player**
G: Is the celebrity you're thinking of alive? J: *Yes.* G: Is the celebrity an actor? J: *Yes.* G: Is the celebrity American? J: *No.* G: Is the celebrity male? J: *Yes.* G: Has the celebrity won an Oscar? J: *No.* G: Is the celebrity known for roles in action films? J: *No.* G: Is the celebrity from the United Kingdom? J: *No.* G: Is the celebrity from Australia? J: *No.* G: Is the celebrity from Asia? J: *No.* G: Is the celebrity from Europe? J: *No.* G: Is the celebrity from Canada? J: *No.* G: Is the celebrity from South America? J: *No.* G: **Is the celebrity from Africa?** J: *Yes.* G: Is the celebrity known for roles in comedic films? J: *Yes.* G: Is the celebrity Trevor Noah? J: *Bingo!*	G: Is the person male? J: *Yes.* G: Is the person on TV? J: *Yes.* G: Does the person host a late night show? J: *Yes.* G: Is the show factual? J: *Dunno.* G: Is the person Stephen Colbert J: *No.* G: Is the person a news anchor? J: *No.* G: Is the person a comedian? J: *Yes.* G: Is the person John Oliver? J: *No.* G: Is the person Trevor Noah? J: *Bingo!*
G: Is your celebrity alive? J: *Yes.* G: Is your celebrity an actor? J: *Yes.* G: Is your celebrity female? J: *No.* G: Has your celebrity won an Oscar? J: *No.* G: Is your celebrity American? J: *No.* G: Is your celebrity British? J: *No.* G: Is your celebrity from Australia? J: *No.* G: Is your celebrity from Asia? J: *No.* G: Is your celebrity from Europe? J: *No.* G: Is your celebrity from Canada? J: *No.* G: Is your celebrity from Africa? J: *Yes.* G: Has your celebrity starred in a superhero movie? J: *No.* G: Is your celebrity over the age of 50? J: *No.* G: **Has your celebrity acted in a film that was nominated for an Academy Award?** J: *Yes.* G: Is your celebrity from South Africa? J: *Yes.* G: Is your celebrity Trevor Noah? J: *Bingo!*	G: Is your celebrity a movie star? J: *Yes.* G: Is the celebrity female? J: *No.* G: Is the celebrity over 50 years old? J: *No.* G: Is the celebrity over 30 years old? J: *Yes.* G: Has this celebrity been in a Marvel movie? J: *No.* G: Is this celebrity known for comedies? J: *Yes.* G: Is this celebrity african american? J: *No.* G: Has this celebrity also been in a TV show? J: *Yes.* G: Has this celebrity been in a movie grossing more than a billion dollars? J: *No.* G: Is this celebrity still making hit movies? J: *No.* G: Was this celebrity famous in the 90s? J: *No.* G: Was this celebrity famous in the 2000s? J: *Yes.* G: Has this celebrity worked with Judd Apatow? J: *No.* G: Was this celebrity in one of the Hangover movies? J: *No.* G: Is this celebrity American? J: *Yes.* G: Was this person on network TV? J: *Yes.* G: Does this celebrity have a famous spouse? J: *No.* G: Is this celebrity over 40 years old? J: *No.* G: Is it Zac Efron? J: *No.* G: Is it Zachary Levi? J: *No.*

Table 8: The gameplay scenario generated using GPT-4 model vs the gameplay scenario by human players, on *Things*. The questions need in-domain knowledge is **highlighted**.

Successful game	Planning issue (Early Enumeration)
G: Is the celebrity alive? *J: Yes.* G: Is the celebrity a male? *J: No.* G: Is she an actress? *J: Yes.* G: Is she American? *J: Yes.* G: Has she won an Academy Award? *J: Yes.* G: Is she above 50 years old? *J: No.* G: Has she starred in a superhero movie? *J: Yes.* G: Is she in the Marvel Cinematic Universe? *J: Yes.* G: Did she play a character that has their own standalone film in the Marvel Cinematic Universe? *J: Yes.* G: Has she played a human with supernatural powers? *J: No.* G: Did she play the character of Black Widow? *J: No.* G: Did she play Captain Marvel? *J: No.* G: Did she play the character of Wasp in the Ant-Man series? *J: Yes.* G: Is the celebrity Evangeline Lilly? *J: No.* G: Did she play the character of Gamora in the Guardians of the Galaxy series? *J: No.* G: Did she play the character of Pepper Potts in the Iron Man series? *J: Yes.* G: Is the celebrity Gwyneth Paltrow? *J: Bingo!*	G: Is the celebrity a person who has been in the public eye for at least 10 years? *J: Yes.* G: Is the celebrity primarily known for acting? *J: Yes.* G: Has the celebrity won an Academy Award? *J: No.* G: Has the celebrity been in any Marvel Cinematic Universe films? *J: Yes.* G: Is the celebrity primarily known for their work in the **Marvel Cinematic Universe**? *J: No.* G: Has the celebrity been in any **DC Extended Universe films**? *J: No.* G: Has the celebrity been in any **Star Wars films**? *J: No.* G: Is the celebrity primarily known for their work in television? *J: No.* G: Has the celebrity been in any reality TV shows? *J: No.* ...
Planning issue (Redundancy)	**Reasoning issue (Inconsistency)**
G: Is the celebrity a musician? *J: No.* G: Is the celebrity primarily known for their acting career? *J: Yes.* G: Is the celebrity a former child actor? *J: No.* G: Is the celebrity primarily known for their work in film? *J: Yes.* G: Is the celebrity mostly known for their work in the past decade? *J: Yes.* G: Is the celebrity a **British national**? *J: No.* G: Is the celebrity primarily known for their work in American cinema? *J: Yes.* G: Is the celebrity a former child star who later became a Hollywood movie star? *J: Yes.* G: Is the celebrity Emma Watson? *J: No.* G: Is the celebrity a **male actor**? *J: No.* G: Is the celebrity a **British female actor**? *J: No.* ...	G: Is the celebrity **male**? *J: Yes.* G: Has the celebrity won an Academy Award? *J: No.* G: Is the celebrity primarily known for their work in film? *J: No.* G: Is the celebrity primarily known for their work in television? *J: No.* G: Is the celebrity known for their work in music? *J: No.* G: Does the celebrity have a significant presence on social media? *J: Yes.* G: Is the celebrity a comedian? *J: No.* G: Is the celebrity a singer? *J: No.* G: Is the celebrity a dancer? *J: No.* ... G: Is the celebrity a political figure? *J: No. You must guess now* G: The celebrity is **Ellen DeGeneres**. Is it right? *J: No.*

Table 9: Demonstration of failure modes of the gameplay with hidden celebrity names. The successful gameplay scenario is generated using GPT-4 model as G. The remaining gameplay scenarios are generated using vicuna-7b model as G.

Artifact: Chain Reaction	Artifact: Self-answering
G: Is the entity a living thing? *J: Yes.* G: Is the living thing an animal? *J: Yes.* G: Is the animal a mammal? *J: Yes.* G: Is the mammal a warm-blooded? *J: Yes.* G: Is the warm-blooded mammal a primate? *J: No.* G: Is the primate a marmoset? *J: No.* G: Is the marmoset a rodent? *J: No.* G: Is the rodent a squirrel? *J: No.* G: Is the squirrel a rabbit? *J: No.*	G: Is the entity you are thinking of a person? Yes *J: No.* G: Is the entity you are thinking of an animal? Yes *J: Yes.* G: Is the animal you are thinking of a mammal? Yes *J: No.* G: Is the animal you are thinking of a bird? Yes *J: No.* G: Is the animal you are thinking of a reptile? Maybe *J: No.* G: Is the animal you are thinking of an amphibian? Maybe *J: No.* G: Is the animal you are thinking of a fish? Yes *J: No.* G: Is the animal you are thinking of a crustacean? Maybe *J: No.* G: Is the animal you are thinking of an arachnid? Maybe *J: No.* G: Is the animal you are thinking of an insect? Yes *J: No.* G: Is the animal you are thinking of a mollusk? No *J: No.* ...

Table 10: Demonstration of failure modes from Vicuna-7b model where the model was trapped by repetitive or problematic patterns.

Celebrity: LeBron James	
Vicuna 7B	**V-FT 7B (Things)**
G: Was this celebrity a singer? *J: No.* G: Was this celebrity a musician? *J: No.* G: Was this celebrity an actor? *J: No.* G: Was this celebrity a comedian? *J: No.* G: Was this celebrity a public figure? *J: Yes.* G: Was this public figure a politician? *J: No.* G: Was this public figure a sports person? *J: Yes.* G: Was this sports person a professional athlete? *J: Yes.* G: Was this professional athlete a tennis player? *J: No.* G: Was this professional athlete a boxer? *J: No.* G: Was this professional athlete a golfer? *J: No.* G: Was this professional athlete a football player? *J: No.* G: Was this professional athlete a basketball player? *J: Yes.* G: Was this basketball player a member of the NBA? *J: Yes.* G: Was this basketball player active during the 1990s? *J: No.* G: Was this basketball player active during the 2000s? *J: Yes.* G: Was this basketball player known for his clutch performances? *J: Yes.* G: Can you name the basketball player? *J: No.* G: I'm sorry, I'll have to ask for a hint. Could it be someone who played for the Los Angeles Lakers during the 2000s? *J: No. You must guess now, what's it?* G: Could it be someone who played for the Los Angeles Clippers during the 2000s? Is it right? *J: No.*	G: Is the celebrity you're thinking of a male? *J: Yes.* G: Is the celebrity you're thinking of a singer? *J: No.* G: Is the celebrity you're thinking of an actor? *J: No.* G: Is the celebrity you're thinking of a sports player? *J: Yes.* G: Is the celebrity you're thinking of a male sports player? *J: Yes.* G: Is the celebrity you're thinking of an Olympic athlete? *J: Yes.* G: Is the celebrity you're thinking of a swimmer? *J: No.* G: Is the celebrity you're thinking of a track and field athlete? *J: No.* G: Is the celebrity you're thinking of a gymnast? *J: No.* G: Is the celebrity you're thinking of a boxer? *J: No.* G: Is the celebrity you're thinking of a soccer player? *J: No.* G: Is the celebrity you're thinking of a basketball player? *J: Yes.* G: Is the celebrity you're thinking of a former NBA player? *J: Yes.* G: Was the celebrity you're thinking of a dominant scorer? *J: Yes.* G: Did the celebrity you're thinking of have a reputation for being a high-flyer? *J: Yes.* G: Is the celebrity you're thinking of known for their acrobatic dunks? *J: Yes.* G: Is the celebrity you're thinking of Michael Jordan? *J: No.* G: Is the celebrity you're thinking of Kobe Bryant? *J: No.* G: Is the celebrity you're thinking of LeBron James? *J: Bingo!*
Celebrity: Brie Larson	
Vicuna 7B	**V-FT 7B (Things)**
G: Is the celebrity you're thinking of a male? *J: No.* G: Does the celebrity you're thinking of have a sister? *J: Yes.* G: Is the sister of your celebrity the same gender as the celebrity? *J: Yes.* G: Is the celebrity you're thinking of a musician? *J: No.* G: Is the celebrity you're thinking of a member of a pop group? *J: No.* G: The celebrity is not a musician, not a member of a pop group, and not a male. Do you know their name? *J: Yes.* G: You have provided the name of the celebrity. Is it a living person? *J: Yes.* G: Your celebrity is a living person who is not a musician, not a member of a pop group, and not a male. Can you provide their profession? *J: Yes.* ...	G: Is the celebrity male? *J: No.* G: Is the celebrity a singer? *J: No.* G: Is the celebrity an actress? *J: Yes.* G: Has the celebrity won an Academy Award? *J: Yes.* G: Has the celebrity won the Academy Award for Best Actress? *J: Yes.* G: Has the celebrity won multiple Academy Awards for acting? *J: No.* G: Did the celebrity win the Academy Award for Best Actress for her role in a biographical film? *J: Yes.* G: Does the celebrity have blonde hair? *J: No.* G: Has the celebrity played a superhero character in a movie? *J: Yes.* G: Is the celebrity over the age of 40? *J: No.* G: Has the celebrity won Academy Awards for portraying real-life individuals on screen? *J: Yes.* G: Is the celebrity Charlize Theron? *J: No.* G: Is the celebrity Cate Blanchett? *J: No.* G: Is the celebrity Jennifer Lawrence? *J: No.* G: Is the celebrity Brie Larson? *J: Bingo!*

Table 11: Comparison of gameplay scenario on *Celebrities* between Vicuna 7B model vs **V-FT 7B (Things)**.

Celebrity to be deducted: LeBron James				
Current top 4 guesses before generating question (Ranked)				**Generated question**
Leonardo DiCaprio	Jennifer Aniston	Will Smith	Angelina Jolie	G: Is the celebrity male? *J: Yes.*
Brad Pitt	Tom Cruise	Leonardo DiCaprio	Dwayne Johnson	G: Is he an actor? *J: No.*
Justin Bieber	Drake	Elon Musk	Cristiano Ronaldo	G: Is he a musician? *J: No.*
Donald Trump	Elon Musk	Cristiano Ronaldo	Richard Branson	G: Is he involved in sports? *J: Yes.*
Michael Jordan	Tiger Woods	LeBron James	Cristiano Ronaldo	G: Is he a retired athlete? *J: No.*
Cristiano Ronaldo	LeBron James	Lionel Messi	Tom Brady	G: Does he play a team sport? *J: Yes.*
Cristiano Ronaldo	Lionel Messi	LeBron James	Tom Brady	G: Does he play American football? *J: No.*
LeBron James	Lionel Messi	Sidney Crosby	Stephen Curry	G: Does he play basketball? *J: Yes.*
LeBron James	Stephen Curry	Kevin Durant	Kawhi Leonard	G: Is he currently playing in the NBA? *J: Yes.*
LeBron James	Stephen Curry	Kevin Durant	Giannis Antetokounmpo	G: Does he play for the Los Angeles Lakers? *J: Yes.*
LeBron James	Anthony Davis	Russell Westbrook	Carmelo Anthony	G: Is he LeBron James? *J: Bingo!*

Table 12: Top 4 guesses made by GPT-4 during a game of guessing LeBron James. Guesses are generated before asking the clarification questions in the same line.

Repetition 1	Repetition 2
G: Is it a living entity? *J: No.*	G: Is it a living entity? *J: No.*
G: Is it man-made? *J: Yes.*	G: Is it man-made? *J: Yes.*
G: Is it used for communication? *J: No.*	G: Is it something used daily? *J: Maybe.*
G: Is it smaller than a bread box? *J: Yes.*	G: Is it electronic? *J: No.*
G: Is it electronic? *J: No.*	G: Is it something you can hold in your hand? *J: Yes.*
G: Is it made of metal? *J: No.*	G: Is it used for writing or drawing? *J: No.*
G: Is it used in daily life? *J: Yes.*	**G: Is it made of plastic?** *J: Yes.*
G: Is it associated with personal hygiene? *J: Maybe.*	G: Is it used in the kitchen? *J: No.*
G: Is it a container of sorts? *J: No.*	G: Is it a toy? *J: No.*
G: Is it used in cooking or eating? *J: No.*	G: Is it used for personal care or hygiene? *J: No.*
G: Is it used for writing or drawing? *J: No.*	G: Is it used for storage? *J: No.*
G: Is it associated with clothing or fashion? *J: No.*	G: Is it a tool or instrument? *J: Yes.*
G: Is it made of glass? *J: No.*	G: Is it associated with a specific profession or hobby? *J: Maybe.*
G: **Is it made of plastic?** *J: Yes.*	G: Is it used for measuring? *J: No.*
G: Is it found in a bathroom? *J: No.*	G: Does it have moving parts? *J: No.*
G: Is it found in a bedroom? *J: Maybe.*	G: Is it used in crafting or DIY projects? *J: Maybe.*
G: Is it a toy? *J: No.*	G: Is it a ruler? *J: No.*
G: Is it used for storage? *J: No.*	G: Is it a magnifying glass? *J: No.*
G: Is it used for cleaning purposes? *J: No. You must guess now, what's it?*	G: Is it a stencil? *J: No. You must guess now, what's it?*
G: Is it a plastic comb? Is it right? *J: No.*	G: Is it a glue stick? Is it right? *J: No.*

Table 13: GPT-4 consistently fails on "Yoga mat" after confirming the entity is made of plastic.

Entity to be deducted: Yoga Mat			
Current top 3 guesses before generating question (Ranked)			Generated question
computer	tree	book	G: Is it a living entity? *J: No.*
Eiffel Tower	Great Pyramid of Giza	Statue of Liberty	G: Is it a living tangible? *J: Yes.*
book	car	computer	G: Is it a man-made? *J: Yes.*
skyscraper	book	smartphone	G: Can it be used in everyday life? *J: Yes.*
smartphone	car	book	G: Is it electronic? *J: No.*
book	chair	table	G: Is it smaller than a bread box? *J: Yes.*
pencil	book	pair of glasses	G: Is it made of metal? *J: No.*
pen	book	wine glass	G: Is it made out of paper? *J: No.*
wooden spoon	glass cup	plastic comb	**G: Is it made of plastic?** *J: Yes.*
plastic water bottle	plastic comb	plastic spoon	G: Is it used for personal hygiene? *J: No.*
plastic spoon	plastic cup	plastic toy	G: Is it used in the kitchen? *J: No.*
plastic pen	plastic ruler	plastic toy	G: Is it used for entertainment or leisure? *J: No.*
plastic pen	plastic comb	plastic ruler	G: Is it used in an office setting? *J: No.*
plastic comb	plastic spoon	plastic toy car	G: Is it related to education or learning? *J: No.*
plastic comb	plastic spoon	plastic toy car	G: Is it associated with storage or organization? *J: Maybe.*
plastic storage box	plastic organizer drawer	plastic file folder	G: Is it found in a bedroom? *J: Maybe.*
plastic box for storage	plastic hanger	plastic drawer organizer	G: Is it associated with clothing or fashion? *J: No.*
plastic box or container	plastic toy bin	plastic shoe rack	G: Is it typically found in a garage or outdoors? *J: No.*
plastic storage box	plastic toy box	plastic jewelry box	G: Is it used to hold or contain smaller items? *J: No.*
plastic cup	plastic ruler	plastic toy	G: Is it a plastic ruler? *J: No.*

Table 14: Top 3 guesses made by GPT-4 guessing entity "Yoga Mat". Following the question in red, all top guesses focus on plastic.