

Appendix

A DATASET COMPOSITION

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

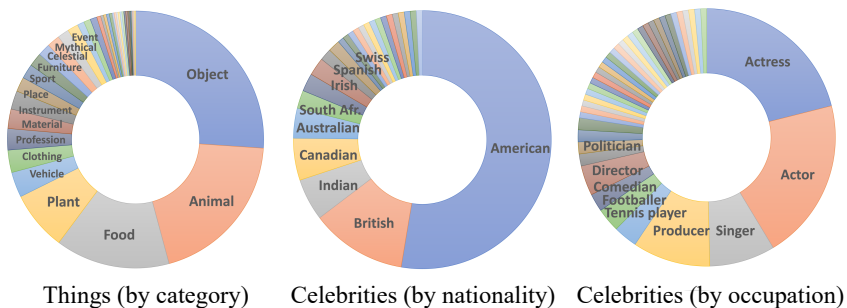


Figure 3: Composition of EDA *Things* and *Celebrities* datasets.

B PROMPT USED FOR *Celebrities*

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}

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 11. 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.

D 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.

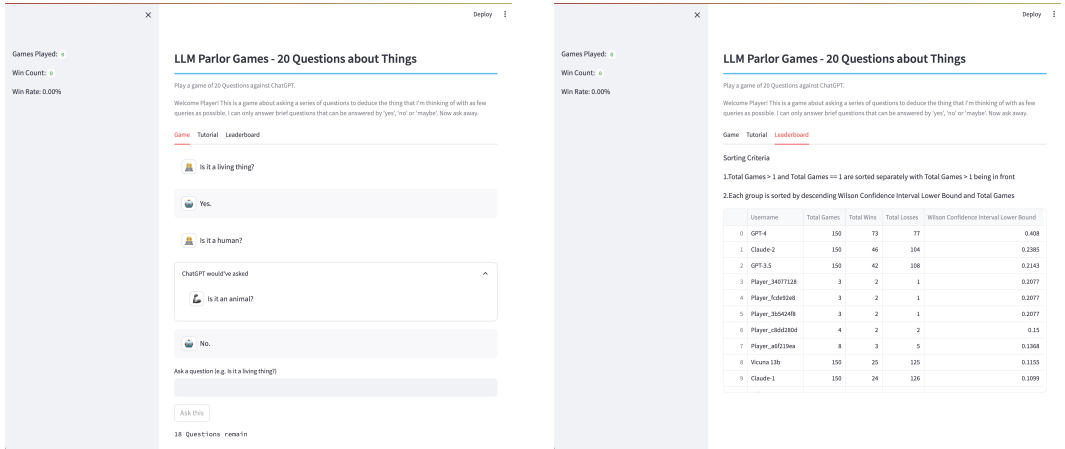


Figure 4: Game play UI interface for collecting human baseline. On the left, human players are given prompt instructions equivalent to LLM guessers. An optional retrospection UI can be toggled to display what ChatGPT would’ve chosen to ask in the last turn. On the right, a leaderboard with Human and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

E COMPARISON OF GPT-4 VS HUMAN PLAYER

We provide additional examples in Table 6 (*Things*) and Table 7 (*Celebrities*) for GPT-4’s generation and compare them with the human players.

E.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 6 (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.

E.2 *Celebrities* DATASET

We present a comparative analysis between GPT-4 and human players on *Celebrities* in Table 7. Our findings reveal that GPT-4 demonstrates a propensity for asking questions that necessitate domain-specific background knowledge (as highlighted in Table 7). 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.

F GAME RUNS AND FAILURE MODES ON *Celebrities* AND *Things*

We provide additional gameplay sessions on *Celebrities*, as shown in Table 8. 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 9. 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.

G EXPERIMENTAL SETTINGS FOR BEHAVIOR 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 \mathcal{J} within the loss function. In all of our experiments, we continued training the model until the validation loss ceased to decrease any further.

H 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 10. 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.

I 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:

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

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.

J FULL ABLATION RESULTS ON RQ2, PLANNING VS REASONING

We provide the full results in Table 5 for this ablation.

	Things				Celebrities			
	#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-4 → Vicuna 7B	16.9±0.2	0.14±0.04	6.1±0.2	0.12±0.03	16.5±0.5	0.23±0.03	7.3±0.1	0.19±0.02
Vicuna 7B	19.1±0.3	0.12±0.05	5.8±0.5	0.10±0.04	19.7±0.3	0.05±0.04	6.2±0.7	0.04±0.03
Vicuna 7B → GPT-4	19.1±0.3	0.14±0.04	5.7±0.2	0.11±0.03	19.7±0.3	0.09±0.04	6.1±0.6	0.06±0.03

Table 5: Planning and reasoning ability comparison between Vicuna 7B and GPT-4. GPT-4 → Vicuna 7B uses GPT-4 to play all except the last turn, swapping in Vicuna 7B in the end to perform reasoning, where as Vicuna 7B → GPT-4 does the opposite.

K 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 12, 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 13. 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
<p>G: Is it a living entity? <i>J: No.</i></p> <p>G: Is it man-made? <i>J: Yes.</i></p> <p>G: Can it be held in one hand? <i>J: Yes.</i></p> <p>G: Is it electronic? <i>J: No.</i></p> <p>G: Is it used for writing or drawing? <i>J: No.</i></p> <p>G: Is it made of metal? <i>J: No.</i></p> <p>G: Is it used in cooking or eating? <i>J: No.</i></p> <p>G: Is it used for entertainment or leisure? <i>J: Yes.</i></p> <p>G: Is it associated with a specific sport? <i>J: No.</i></p> <p>G: Is it used in board games? <i>J: No.</i></p> <p>G: Is it associated with music? <i>J: Yes.</i></p> <p>G: Is it a musical instrument? <i>J: Yes.</i></p> <p>G: Is it a stringed instrument? <i>J: Yes.</i></p> <p>G: Is it commonly used in a symphony orchestra? <i>J: No.</i></p> <p>G: Is it a guitar? <i>J: Bingo!</i></p>	<p>G: Is it a living thing? <i>J: No.</i></p> <p>G: Is it a man-made object? <i>J: Yes.</i></p> <p>G: Is it commonly used indoors? <i>J: Yes.</i></p> <p>G: Is it larger than a toaster? <i>J: Yes.</i></p> <p>G: Is it used for recreational purposes? <i>J: Yes.</i></p> <p>G: Does it require electricity? <i>J: Maybe.</i></p> <p>G: Does it have a screen? <i>J: No.</i></p> <p>G: Does it contain metal? <i>J: Yes.</i></p> <p>G: Does it contain glass? <i>J: No.</i></p> <p>G: Does it make sound? <i>J: Yes.</i></p> <p>G: Is it used to play music? <i>J: Yes.</i></p> <p>G: Is it a speaker? <i>J: No.</i></p> <p>G: Is it a radio? <i>J: No.</i></p> <p>G: Does it have a wireless connection? <i>J: No.</i></p> <p>G: Can it display images? <i>J: No.</i></p> <p>G: Is it some kind of music player? <i>J: No.</i></p> <p>G: Is it a musical instrument? <i>J: Yes.</i></p> <p>G: Is it a guitar? <i>J: Bingo!</i></p>	<p>G: Is it a living thing? <i>J: No.</i></p> <p>G: Is it a man-made object? <i>J: Yes.</i></p> <p>G: Is it commonly used indoors? <i>J: Yes.</i></p> <p>G: Is it larger than a toaster? <i>J: Yes.</i></p> <p>G: Is it used for recreational purposes? <i>J: Yes.</i></p> <p>G: Does it require electricity? <i>J: Maybe.</i></p> <p>G: Does it have a screen? <i>J: No.</i></p> <p>G: Does it contain metal? <i>J: Yes.</i></p> <p>G: Does it contain glass? <i>J: No.</i></p> <p>G: Does it make sound? <i>J: Yes.</i></p> <p>G: Is it used to play music? <i>J: Yes.</i></p> <p>G: Is it a speaker? <i>J: No.</i></p> <p>G: Is it a radio? <i>J: No.</i></p> <p>G: Does it have a wireless connection? <i>J: No.</i></p> <p>G: Can it display images? <i>J: No.</i></p> <p>GPT: Is it a musical instrument? <i>J: Yes.</i></p> <p>GPT: Is it a string instrument? <i>J: Yes.</i></p> <p>GPT: Is it a guitar? <i>J: Bingo!</i></p>
Item: Bald eagle		
GPT-4	Human player	
<p>G: Is it a living thing? <i>J: Yes.</i></p> <p>G: Is it an animal? <i>J: Yes.</i></p> <p>G: Is it a domestic animal? <i>J: No.</i></p> <p>G: Is it a mammal? <i>J: No.</i></p> <p>G: Is it a reptile? <i>J: No.</i></p> <p>G: Is it a bird? <i>J: Yes.</i></p> <p>G: Is it a bird of prey? <i>J: Yes.</i></p> <p>G: Does it have a predominantly white color? <i>J: Yes.</i></p> <p>G: Is it found in the Polar regions? <i>J: No.</i></p> <p>G: Is it a bald eagle? <i>J: Bingo!</i></p>	<p>G: Is it alive? <i>J: Yes.</i></p> <p>G: Does it live on land? <i>J: Yes.</i></p> <p>G: Is it a plant? <i>J: No.</i></p> <p>G: Is it human? <i>J: No.</i></p> <p>G: Does it live in the Americas? <i>J: Yes.</i></p> <p>G: Does it live in North America? <i>J: Yes.</i></p> <p>G: Is it a predator? <i>J: Yes.</i></p> <p>G: Is it a bear? <i>J: No.</i></p> <p>G: Is it a bird? <i>J: Yes.</i></p> <p>G: Is it a bald eagle? <i>J: Bingo!</i></p>	
Item: Jigsaw Puzzle		
GPT-4	Human player	
<p>G: Is it a living entity? <i>J: No.</i></p> <p>G: Is it man-made? <i>J: Yes.</i></p> <p>G: Is it used for communication? <i>J: No.</i></p> <p>G: Is it larger than a microwave? <i>J: Maybe.</i></p> <p>G: Is it used for transportation? <i>J: No.</i></p> <p>G: Is it made of metal? <i>J: No.</i></p> <p>G: Is it used in homes? <i>J: Yes.</i></p> <p>G: Is it used for sitting? <i>J: No.</i></p> <p>G: Is it used for storage? <i>J: No.</i></p> <p>G: Is it a decorative item? <i>J: Maybe.</i></p> <p>G: Is it used in the kitchen? <i>J: No.</i></p> <p>G: Is it used in the living room? <i>J: Maybe.</i></p> <p>G: Is it used for entertainment purposes? <i>J: Yes.</i></p> <p>G: Is it a musical instrument? <i>J: No.</i></p> <p>G: Is it a television? <i>J: No.</i></p> <p>G: Is it a game or a toy? <i>J: Yes.</i></p> <p>G: Is it a board game? <i>J: No.</i></p> <p>G: Is it a video game console? <i>J: No.</i></p> <p>G: Is it a physical toy? <i>J: Yes. You must guess now, what's it?</i></p> <p>G: Is it a puzzle? Is it right? <i>J: Yes.</i></p>	<p>G: What is the meaning of life? <i>J: No.</i></p> <p>G: What is a good thing? <i>J: Maybe.</i></p> <p>G: Is it human? <i>J: No.</i></p> <p>G: Is it an object? <i>J: Yes.</i></p> <p>G: Is it a tool? <i>J: Yes.</i></p> <p>G: Is it a hammer? <i>J: No.</i></p> <p>G: Is it a kitchen tool? <i>J: No.</i></p> <p>G: What is it? <i>J: Maybe.</i></p> <p>G: Does it cut? <i>J: No.</i></p> <p>G: Is it for food? <i>J: No.</i></p> <p>G: Is it for cars? <i>J: No.</i></p> <p>G: Is it big? <i>J: Maybe.</i></p> <p>G: Can you eat it? <i>J: No.</i></p> <p>G: Is it red? <i>J: No.</i></p> <p>G: Is it metal? <i>J: No.</i></p> <p>G: Is it used for communication? <i>J: No.</i></p> <p>G: Is it used for storing? <i>J: No.</i></p> <p>G: Is it used for helping humans? <i>J: Maybe.</i></p> <p>G: Is it used for taking pictures? <i>J: No.</i></p> <p>G: Alarm clock <i>J: No.</i></p>	

Table 6: The gameplay scenario generated using GPT-4 model vs the gameplay scenario by human players, on *Things*.

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

Table 7: 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)
<p>G: Is the celebrity alive? <i>J: Yes.</i></p> <p>G: Is the celebrity a male? <i>J: No.</i></p> <p>G: Is she an actress? <i>J: Yes.</i></p> <p>G: Is she American? <i>J: Yes.</i></p> <p>G: Has she won an Academy Award? <i>J: Yes.</i></p> <p>G: Is she above 50 years old? <i>J: No.</i></p> <p>G: Has she starred in a superhero movie? <i>J: Yes.</i></p> <p>G: Is she in the Marvel Cinematic Universe? <i>J: Yes.</i></p> <p>G: Did she play a character that has their own standalone film in the Marvel Cinematic Universe? <i>J: Yes.</i></p> <p>G: Has she played a human with supernatural powers? <i>J: No.</i></p> <p>G: Did she play the character of Black Widow? <i>J: No.</i></p> <p>G: Did she play Captain Marvel? <i>J: No.</i></p> <p>G: Did she play the character of Wasp in the Ant-Man series? <i>J: Yes.</i></p> <p>G: Is the celebrity Evangeline Lilly? <i>J: No.</i></p> <p>G: Did she play the character of Gamora in the Guardians of the Galaxy series? <i>J: No.</i></p> <p>G: Did she play the character of Pepper Potts in the Iron Man series? <i>J: Yes.</i></p> <p>G: Is the celebrity Gwyneth Paltrow? <i>J: Bingo!</i></p>	<p>G: Is the celebrity a person who has been in the public eye for at least 10 years? <i>J: Yes.</i></p> <p>G: Is the celebrity primarily known for acting? <i>J: Yes.</i></p> <p>G: Has the celebrity won an Academy Award? <i>J: No.</i></p> <p>G: Has the celebrity been in any Marvel Cinematic Universe films? <i>J: Yes.</i></p> <p>G: Is the celebrity primarily known for their work in the Marvel Cinematic Universe? <i>J: No.</i></p> <p>G: Has the celebrity been in any DC Extended Universe films? <i>J: No.</i></p> <p>G: Has the celebrity been in any Star Wars films? <i>J: No.</i></p> <p>G: Is the celebrity primarily known for their work in television? <i>J: No.</i></p> <p>G: Has the celebrity been in any reality TV shows? <i>J: No.</i></p> <p>...</p>
Planning issue (Redundancy)	Reasoning issue (Inconsistency)
<p>G: Is the celebrity a musician? <i>J: No.</i></p> <p>G: Is the celebrity primarily known for their acting career? <i>J: Yes.</i></p> <p>G: Is the celebrity a former child actor? <i>J: No.</i></p> <p>G: Is the celebrity primarily known for their work in film? <i>J: Yes.</i></p> <p>G: Is the celebrity mostly known for their work in the past decade? <i>J: Yes.</i></p> <p>G: Is the celebrity a British national? <i>J: No.</i></p> <p>G: Is the celebrity primarily known for their work in American cinema? <i>J: Yes.</i></p> <p>G: Is the celebrity a former child star who later became a Hollywood movie star? <i>J: Yes.</i></p> <p>G: Is the celebrity Emma Watson? <i>J: No.</i></p> <p>G: Is the celebrity a male actor? <i>J: No.</i></p> <p>G: Is the celebrity a British female actor? <i>J: No.</i></p> <p>...</p>	<p>G: Is the celebrity male? <i>J: Yes.</i></p> <p>G: Has the celebrity won an Academy Award? <i>J: No.</i></p> <p>G: Is the celebrity primarily known for their work in film? <i>J: No.</i></p> <p>G: Is the celebrity primarily known for their work in television? <i>J: No.</i></p> <p>G: Is the celebrity known for their work in music? <i>J: No.</i></p> <p>G: Does the celebrity have a significant presence on social media? <i>J: Yes.</i></p> <p>G: Is the celebrity a comedian? <i>J: No.</i></p> <p>G: Is the celebrity a singer? <i>J: No.</i></p> <p>G: Is the celebrity a dancer? <i>J: No.</i></p> <p>...</p> <p>G: Is the celebrity a political figure? <i>J: No. You must guess now</i></p> <p>G: The celebrity is Ellen DeGeneres. Is it right? <i>J: No.</i></p>

Table 8: 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
<p>G: Is the entity a living thing? <i>J: Yes.</i></p> <p>G: Is the living thing an animal? <i>J: Yes.</i></p> <p>G: Is the animal a mammal? <i>J: Yes.</i></p> <p>G: Is the mammal warm-blooded? <i>J: Yes.</i></p> <p>G: Is the warm-blooded mammal a primate? <i>J: No.</i></p> <p>G: Is the primate a marmoset? <i>J: No.</i></p> <p>G: Is the marmoset a rodent? <i>J: No.</i></p> <p>G: Is the rodent a squirrel? <i>J: No.</i></p> <p>G: Is the squirrel a rabbit? <i>J: No.</i></p> <p>...</p>	<p>G: Is the entity you are thinking of a person? Yes <i>J: No.</i></p> <p>G: Is the entity you are thinking of an animal? Yes <i>J: Yes.</i></p> <p>G: Is the animal you are thinking of a mammal? Yes <i>J: No.</i></p> <p>G: Is the animal you are thinking of a bird? Yes <i>J: No.</i></p> <p>G: Is the animal you are thinking of a reptile? Maybe <i>J: No.</i></p> <p>G: Is the animal you are thinking of an amphibian? Maybe <i>J: No.</i></p> <p>G: Is the animal you are thinking of a fish? Yes <i>J: No.</i></p> <p>G: Is the animal you are thinking of a crustacean? Maybe <i>J: No.</i></p> <p>G: Is the animal you are thinking of an arachnid? Maybe <i>J: No.</i></p> <p>G: Is the animal you are thinking of an insect? Yes <i>J: No.</i></p> <p>G: Is the animal you are thinking of a mollusk? No <i>J: No.</i></p> <p>...</p>

Table 9: 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)
<p>G: Was this celebrity a singer? <i>J: No.</i></p> <p>G: Was this celebrity a musician? <i>J: No.</i></p> <p>G: Was this celebrity an actor? <i>J: No.</i></p> <p>G: Was this celebrity a comedian? <i>J: No.</i></p> <p>G: Was this celebrity a public figure? <i>J: Yes.</i></p> <p>G: Was this public figure a politician? <i>J: No.</i></p> <p>G: Was this public figure a sports person? <i>J: Yes.</i></p> <p>G: Was this sports person a professional athlete? <i>J: Yes.</i></p> <p>G: Was this professional athlete a tennis player? <i>J: No.</i></p> <p>G: Was this professional athlete a boxer? <i>J: No.</i></p> <p>G: Was this professional athlete a golfer? <i>J: No.</i></p> <p>G: Was this professional athlete a football player? <i>J: No.</i></p> <p>G: Was this professional athlete a basketball player? <i>J: Yes.</i></p> <p>G: Was this basketball player a member of the NBA? <i>J: Yes.</i></p> <p>G: Was this basketball player active during the 1990s? <i>J: No.</i></p> <p>G: Was this basketball player active during the 2000s? <i>J: Yes.</i></p> <p>G: Was this basketball player known for his clutch performances? <i>J: Yes.</i></p> <p>G: Can you name the basketball player? <i>J: No.</i></p> <p>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? <i>J: No. You must guess now, what's it?</i></p> <p>G: Could it be someone who played for the Los Angeles Clippers during the 2000s? Is it right? <i>J: No.</i></p>	<p>G: Is the celebrity you're thinking of a male? <i>J: Yes.</i></p> <p>G: Is the celebrity you're thinking of a singer? <i>J: No.</i></p> <p>G: Is the celebrity you're thinking of an actor? <i>J: No.</i></p> <p>G: Is the celebrity you're thinking of a sports player? <i>J: Yes.</i></p> <p>G: Is the celebrity you're thinking of a male sports player? <i>J: Yes.</i></p> <p>G: Is the celebrity you're thinking of an Olympic athlete? <i>J: Yes.</i></p> <p>G: Is the celebrity you're thinking of a swimmer? <i>J: No.</i></p> <p>G: Is the celebrity you're thinking of a track and field athlete? <i>J: No.</i></p> <p>G: Is the celebrity you're thinking of a gymnast? <i>J: No.</i></p> <p>G: Is the celebrity you're thinking of a boxer? <i>J: No.</i></p> <p>G: Is the celebrity you're thinking of a soccer player? <i>J: No.</i></p> <p>G: Is the celebrity you're thinking of a basketball player? <i>J: Yes.</i></p> <p>G: Is the celebrity you're thinking of a former NBA player? <i>J: Yes.</i></p> <p>G: Was the celebrity you're thinking of a dominant scorer? <i>J: Yes.</i></p> <p>G: Did the celebrity you're thinking of have a reputation for being a high-flyer? <i>J: Yes.</i></p> <p>G: Is the celebrity you're thinking of known for their acrobatic dunks? <i>J: Yes.</i></p> <p>G: Is the celebrity you're thinking of Michael Jordan? <i>J: No.</i></p> <p>G: Is the celebrity you're thinking of Kobe Bryant? <i>J: No.</i></p> <p>G: Is the celebrity you're thinking of LeBron James? <i>J: Bingo!</i></p>
Celebrity: Brie Larson	
Vicuna 7B	V-FT 7B (Things)
<p>G: Is the celebrity you're thinking of a male? <i>J: No.</i></p> <p>G: Does the celebrity you're thinking of have a sister? <i>J: Yes.</i></p> <p>G: Is the sister of your celebrity the same gender as the celebrity? <i>J: Yes.</i></p> <p>G: Is the celebrity you're thinking of a musician? <i>J: No.</i></p> <p>G: Is the celebrity you're thinking of a member of a pop group? <i>J: No.</i></p> <p>G: The celebrity is not a musician, not a member of a pop group, and not a male. Do you know their name? <i>J: Yes.</i></p> <p>G: You have provided the name of the celebrity. Is it a living person? <i>J: Yes.</i></p> <p>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? <i>J: Yes.</i></p> <p>...</p>	<p>G: Is the celebrity male? <i>J: No.</i></p> <p>G: Is the celebrity a singer? <i>J: No.</i></p> <p>G: Is the celebrity an actress? <i>J: Yes.</i></p> <p>G: Has the celebrity won an Academy Award? <i>J: Yes.</i></p> <p>G: Has the celebrity won the Academy Award for Best Actress? <i>J: Yes.</i></p> <p>G: Has the celebrity won multiple Academy Awards for acting? <i>J: No.</i></p> <p>G: Did the celebrity win the Academy Award for Best Actress for her role in a biographical film? <i>J: Yes.</i></p> <p>G: Does the celebrity have blonde hair? <i>J: No.</i></p> <p>G: Has the celebrity played a superhero character in a movie? <i>J: Yes.</i></p> <p>G: Is the celebrity over the age of 40? <i>J: No.</i></p> <p>G: Has the celebrity won Academy Awards for portraying real-life individuals on screen? <i>J: Yes.</i></p> <p>G: Is the celebrity Charlize Theron? <i>J: No.</i></p> <p>G: Is the celebrity Cate Blanchett? <i>J: No.</i></p> <p>G: Is the celebrity Jennifer Lawrence? <i>J: No.</i></p> <p>G: Is the celebrity Brie Larson? <i>J: Bingo!</i></p>

Table 10: Comparison of gameplay scenario on *Celebrities* between Vicuna 7B model vs V-FT 7B (Things).

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

Table 11: 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. Striked guesses are eliminated after the question is answered in a retrospective perspective.

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

Table 12: GPT-4 consistently fails on “Yoga mat” after confirming the entity is made of plastic.

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

Table 13: Top 3 guesses made by GPT-4 guessing entity “Yoga Mat”. Following the question in red, all top guesses focus on plastic.