ChatGPT's Information Seeking Strategy: Insights from the 20-Questions Game

Leonardo Bertolazzi
University of Trento
leonardo.bertolazzi@unitn.it

Davide Mazzaccara
University of Trento
davide.mazzaccara@unitn.it

Filippo Merlo
University of Trento
filippo.merlo@studenti.unitn.it

Raffaella Bernardi
University of Trento
raffaella.bernardi@unitn.it

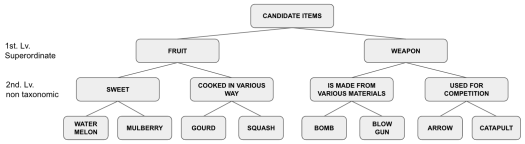
Abstract

Large Language Models, and ChatGPT in particular, have recently grabbed the attention of the community and the media. Having reached high language proficiency, attention has been shifting toward its reasoning capabilities. In this paper, our main aim is to evaluate ChatGPT's question generation in a task where language production should be driven by an implicit reasoning process. To this end, we employ the 20-Questions game, traditionally used within the Cognitive Science community to inspect the information seeking-strategy's development. This task requires a series of interconnected skills: asking informative questions, stepwise updating the hypothesis space, and stopping asking questions when enough information has been collected. We build hierarchical hypothesis spaces, exploiting feature norms collected from humans vs. ChatGPT itself, and we inspect the efficiency and informativeness of ChatGPT's strategy. Our results show that ChatGPT's performance gets closer to an optimal agent only when prompted to explicitly list the updated space stepwise.

1 Introduction

ChatGPT's impressive ability to solve numerous natural language tasks has put it in the spotlight of Academia and media attention (Bang et al., 2023; Laskar et al., 2023). The success on a variety of tasks has brought people to even claim that GPT-4 "could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI)" (Bubeck et al., 2023). Others are more cautious, showing the weakness of the model's reasoning abilities, (e.g., Bang et al. 2023).

A core aspect of human intelligence is the implicit connection between the reasoning process and language production. This connection strongly drives the generation of questions in information-seeking scenarios which, therefore, have been largely studied in Cognitive Science. After the pioneering work by Mosher and Hornsby (1966), the 20-Questions game has been employed to observe children's cognitive developmental trajectory: A player thinks of an entity, the second player is given a set of candidates (e.g., *cat*, *dog*, *bird*) and has to identify the target entity among the possible candidates by making Yes-or-No questions. This and following experiments have shown that through the developmental trajectory, children learn to recognize object-general features, cluster similar objects into categories and use such categorization to ask context dependent informative questions: they shift from *Hypothesis Scanning* questions ("*Is it a dog?*") to *Constraint-Seeking* questions ("*Does it has four legs?*"). Such a shift let the elder children be more efficient in their information seeking process. Moreover, pre-scholar children tend to continue asking questions when enough information has been collected (i.e., the space has reduced to one candidate). They do not know when to stop (Ruggeri et al., 2016), a core skill of infor-



Questioner	Answerer	Sp
1. Is it a weapon?	No	4
2. Is it sweet?	Yes	2
3. Is it the watermelon?	No	1
4. Is it the mulberry?	Yes	

Figure 1: **Upper part**: an example of a Hierarchical Hypothesis Space built with ChatGPT-feature norms. **Bottom part**: an example of an optimal Questioner which always divide the space (Sp) into half (starting with 8 candidates, going to 4:4, and then 2:2). Questions at turns 1-2 are constraint-seeking (CS), while the 3rd and 4th are hypothesis-scanning (HS). With the halfsplit procedure, the target can be guessed with just 3 turns or at most with 4, as in the example.

Proceedings of the 16th International Natural Language Generation Conference, pages 153–162
September 11–15, 2023.

mation search and decision-making (Todd et al., 1999). Interestingly, Ruggeri et al. (2021) uses a hierarchical version of the 20-Q game, in which candidates are organized into three category levels based on shared features; by providing children with the object-related features needed to halfsplit the space, children were able to target such higher category levels, reaching the solution more efficiently. Inspired by this literature, we use a hierarchical 20-Q game to evaluate whether ChatGPT is able to generate questions driven by its reasoning over the Hypothesis Space (HypSp).

We leverage on the widely used feature norms elicitated from human annotators, McRae-norms (McRae et al., 2005) to build the hierarchical hypothesis spaces. Such norms reflect humans' knowledge representation which could differ from ChatGPT's knowledge. To mitigate this potential difference, we build also hypothesis spaces using norms elicitated from ChatGPT itself, (GPT-norms (Hansen and Hebart, 2022)). Figure 1 reports an example of an 8 candidate symmetric nested space based on GPT-norms.

We prompt ChatGPT[1] to play the 20-Q game, both in the role of the Questioner and of the Answerer.[2] We aim to understand whether the (a) Questioner is able to identify the high-level property that clusters the space and hence asks whether the target has that property; (b) it also knows whether all the other candidates have or do not have that property, and is able to use such information to update the HypSp stepwise; and finally (c) it understands when to stop asking questions, i.e., the HypSp is reduced to a singleton. Figure 1 includes a dialogue an optimal agent could ask, if driven by an half-split search. Our results show that ChatGPT's performance is far from an optimal agent when having to update the space internally and it is closer to it when prompted to explicitly list the updated space stepwise.

2 Related Work

Our work put together two research lines: the current effort of the AI community to evaluate ChatGPT language and reasoning skills, and the cognitive science literature focusing on the developmental trajectory of information search strategies in humans.

ChatGPT evaluation Bang et al. (2023) run a deep and broad evaluation of ChatGPT on a variety of well recognized benchmarks in the Natural Language Processing community. ChatGPT results to be State-of-the-Art in zero-shot setting for most natural language understanding tasks. Though it is more suitable for open-domain dialogue tasks, it performs well also in task-oriented dialogues, and it is able to keep track of information given in previous turns, when answering follow up questions. Moreover, Bang et al. (2023) evaluate ChatGPT reasoning skills: though it lacks inducting reasoning skills, it performs well on deductive reasoning in clean settings. However, as other LLMs (Ott et al., 2023), ChatGPT as well encounters problems with complex deductive reasoning involving multi-hops, viz. a combination of facts spread in different passages of a corpus. Zhu et al. (2023) challenged ChatGPT on the Visual Dialogue task, originally proposed by Das et al. (2017). The informativeness of the question is measured on the quality of the caption it summarises out of the dialogue. As far as we know, this is the first work to evaluate the information seeking strategy of a LLM using the 20 Questions game. Our research question is whether and to what extent the language generation of a LLM is tied to reasoning.

Developmental and Cognitive Psychology Starting with Mosher and Hornsby (1966), the 20-Q search task has been largely used in developmental and cognitive psychology. Among the measures to evaluate the question's informativeness, Expected Information Gain (Lindley, 1956) emerges as one of the most used. It values questions with respect to the uncertainty reduction, and it is usually connected with the prior probability. Subjects have been evaluated with the 20-Q game considering both scenarios simulating *prior* expectations and scenarios with uniform distribution (e.g., Ruggeri and Lombrozo 2015; Meder et al. 2019; Ruggeri et al. 2021; Testoni 2023). Our scenario is the uniform distribution.

It is widely accepted that children's search strategies are less efficient than adults' ones. Rather than identifying high-order properties splitting efficiently the Hypothesis Space, indeed, children tend to scan the space item by item. In complex scenarios, it has been shown that adults do not efficiently plan ahead; they tend to follow a half-splitting strategy: ask the question that more closely approximates a division of the space into half (Meder

[1] We used the API version of `gpt-3.5-turbo` available between March and May.

[2] The data and scripts associated to this paper are available at https://github.com/leobertolazzi/20q-chatgpt.

et al., 2019). Rothe et al. (2018) show that people can accurately evaluate questions quality, but have limited ability to optimize the informativeness of their questions. By leverage of feature norm collections, we work on a simplified scenario where adults would more easily stay close to an optimal agent.

Feature Norms Feature norms refer to minimal semantic descriptions that capture the typical attributes associated with a collection of objects or concepts (e.g., a *dog* can be described by features such as *has fur* and *does bark*). One common method of acquiring semantic features for concepts is to ask individuals to list properties associated with a given concept. A broadly used collection is the McRae-norms (McRae et al., 2005) which comprise 2524 unique features collected from approximately 725 participants, which are in turn categorized according to Wu and Barsalou (2009)'s taxonomy of relations (WB). These norms encompass 541 animate and inanimate concrete concepts, with an average of 30 participants providing feature listings for each of them. The features included in the McRae-norms are of various types, such as physical (perceptual) properties, functional properties, taxonomic properties, and encyclopedic facts. Inspired by this work and to obtain a large-scale collection, (Hansen and Hebart, 2022) instructed GPT-3 (Brown et al., 2020) to generate semantic feature norms for a diverse set of 1,854 concrete concepts which have been annotated with 84561 unique features elicitated from the model through 30 runs, pre-processed and filtered. These feature norms were then released by the authors for public use and exploration; the authors expanded their method to other models of the GPT family when they became available. We use the feature norms obtained from GPT-3.5-turbo, and refer to them as GPT-norms.[3]

3 Hierarchical version of the 20-Q game

Following Ruggeri et al. (2021), we created a hierarchical version of the 20-Q game. In other words, the hypothesis spaces are built out of two subsets of equal size (N:N), and iteratively divided into further subsets based on some other features. We exploit McRae (McRae et al., 2005) and ChatGPT (Hansen and Hebart, 2022) feature norms to build the nested structures. We consider hierarchies of two levels (8 candidates, divided into 4:4, and 2:2) and of three levels (16 candidates, divided into 8:8, 4:4, and 2:2). The first level is always based on superordinate properties ($F_{1_a} vs.\ F_{1b}$), which are by definition mutually exclusive (e.g., *bird* vs. *mammal*, *fruit* vs. *weapon*, etc.). The subsets of the other levels instead are obtained from all the other feature norms associated with the candidates (e.g., items that are *fruit* could be divided into those that are *sweet* vs. those that are *cooked in various way*). We make sure that the feature that is shared by half of the candidates is not listed for any of the item in the other half, and viceversa. The leaves of the hierarchy are randomly selected among the concepts of the corresponding groups. See Figure 1 for the schema and an example with 8 candidates organized based on GPT-norms.

Hierarchical Hypothesis Space creation Our starting point are the concepts in McRae et al. (2005) and in Hansen and Hebart (2022), 541 and 1854, respectively. For McRae-norms, we selected the superordinates frequent enough to let us create spaces of 8 and of 16 candidates.[4] For the second level, we use features of the other 8 most frequent WB relations (51 unique features). For the Hypothesis Spaces built from ChatGPT-norms, we selected the same 6 superordinates for the first level splits, and other 806 most frequent unique features for the second levels. We built the hypothesis space through a recursive process that guarantees variety and randomness of the selection (See the Supplementary Material for details). We will refer to these two types of hypothesis spaces as 8 vs. 16 candidates sets (cds), distinguishing the former into McRae- and GPT-based 8 cds; henceforth, `8-McRae`, `16-McRae`, `8-GPT`.

Game creation A game consists of a set of candidates, assigned to the Questioner player, and a target among them, assigned to the Answerer. We build 90 games for each of the three types of Hypothesis Space as follows. First of all, out of the 6 selected superordinate features, we build all possible pairs, viz. 15 (F_{1a}, F_{1b}); we then randomly select 6 sets of candidates for each of the 15 pairs, yielding 90 unique sets (total 270 sets). Finally, we build the 90 games by randomly selecting the target 3 times from the candidates that share F_{1a}

[3] The norms collected with `gpt-3.5-turbo` are available at https://github.com/ViCCo-Group/semantic_features_gpt_3.

[4] The 6 superordinates we use to build the first level splits are: mammal, bird, clothing, weapon, fruit, vegetable.

and 3 times from those that share F_{1b}. This process guarantees variety of the concepts and targets.

4 Agent Roles

Below we describe how we employ ChatGPT as game players to generate the dialogues and as diagnostic agents to evaluate the Questioner's information seeking ability.

4.1 Game Players

To generate the dialogues, ChatGPT is instructed to play the role of the Questioner (**ChatGPT-Q**) and of the Answerer (**ChatGPT-A**) with a similar system prompts. The shared part of the prompts explicitly states the only possible answers are 'yes'and 'no'. ChatGPT-Q is told to ask as few questions as possible; the Questioner starts by asking the first question, which is appended to the Answerer's prompt in order to generate the first answer. In this way, the dialogue history is iteratively increased after each turn. ChatGPT-A is told to acknowledge when the Questioner has correctly guessed the item by answering "Yes! That's correct.". Focusing on ChatGPT's capabilities of reasoning about the hypothesis space and asking questions that reflect such reasoning, we retain only successful dialogues. More precisely, the dialogue is kept if the Answerer considers the target reached. Our evaluation is focused on the Questioner role, hence, for it we define theoretically an upper and a lower bound as described below.

We take as upper-bound a model that similarly to adults seeks for a property shared by several items in the space. In particular, we use the **optimal agent** which acts similarly to a binary search algorithm: at each turn, it divides in half the hypothesis space under discussion ($N/2$). When only two items are left, the optimal agent makes a guess that has the 50% chance of being the correct target. This half-split strategy takes on average $\log_2 N + 1/2$ turns to solve the game, where N is the number of items at the beginning of the game.

As lower-bound we consider a model close to the 4-Y child who tends to scan the space item by item. Therefore, our **baseline agent** acts similarly to a linear search algorithm: at each turn, it divides the space into $1\ vs.\ N-1$. Given N items at the beginning of the game, it takes on average $N/2$ turns to solve the game.

4.2 Diagnostic Agents

To evaluate the model's ability to stepwise reduce the hypothesis space we exploit ChatGPT in the role of an external Oracle (ChatGPT-Oracle), and of an external Guesser (ChatGPT-Guesser). Moreover, we activate the Guesser internal to the Questioner by prompting the model to update the candidates at each turn (ChatGPT-Q-stepwise).

ChatGPT-Oracle is given a question in the dialogue sets described in the previous section and for each item in the hypothesis space of the corresponding game says whether the item has or does not have the required property.[5] This provides us with Y/N-annotation of the hypothesis space that we use to obtain a "ground truth" updated space at each turn. The feasibility of such method relies on the fact that the dialogues are rather simple and no actual linguistic dependencies are in place between the turns (See Supplementary Material for details). **ChatGPT-Guesser** is given chucks of the dialogue histories generated by the game players and is asked to list the candidates till the given turn. Finally, we modify the prompt of ChatGPT-Q by asking it to list the candidates under discussion stepwise before asking a new question (**ChatGPT-Q-stepwise**). The prompts used for each role are reported in the Supplementary Material.

5 Experimental Setup

We are interested in understanding whether ChatGPT's language generation is driven by its reasoning process. To answer this question, we propose a number of measures aimed to shed light on the reasoning processes that are implicit in the game: identify the high-level property shared by several items, update the space stepwise, and efficiently arrive to a space with just one possible candidate and realize that it is time to stop asking questions. Not having the possibility to run an ablation study of the model, we simulate it by comparing ChatGPT-Q, simply prompted to play the game, with ChatGPT-Q-stepwise which is explicitly told to update the space turn by turn.

Information seeking strategy Following the method used in the Cognitive Science literature to evaluate children's developmental phases, we evaluated the information seeking strategy used

[5] We verified the reliability of the annotation by evaluating the model's accuracy on a sample of 180 questions: it correctly answered 83% of the questions.

by ChatGPT-Q by observing the type of questions it asks and their informativeness. First of all, we compute the percentage of questions that are *Hypothesis-Scanning (HS)* and *Constraint-Seeking (CS)*. A question is considered *HS* iff it explicitly mentions one of the candidates in the hypothesis space. All the other questions are considered *CS*. We compute the percentage of HS and CS questions within a game and by the position of the turn within the dialogue.[6]

Following Ruggeri et al. (2016), Meder et al. (2019) and Testoni (2023), we compute the *Expected Information Gain (EIG)* of each question and report the average EIG per turn.[7] As clearly explained in Meder et al. (2019), the information gain of a question is the entropy in the space (given by the number of items and the associated probability) at turn t_i before asking the question minus the expected entropy after asking it (t_{i+1}):

$$IG = H_{t_i} - H_{t_{i+1}}$$

As in Meder et al. (2019), in our case, we consider all items in the space to be equally likely to be the target. Hence, what defines entropy is the number of items within the subsets answered with Yes vs. No, based on the external Oracle annotation.

A model that asks fewer HS, especially in the earlier turns is closer to the more efficient strategy used by adults. Its question EIG is expected to be very high in the early turns and to decrease in the later ones.

Hypothesis space update A core skill of the Questioner playing the 20-Q game is the ability to *mentally* keep the space of the hypothesis updated stepwise. We evaluate whether ChatGPT-Q is able to update at each turn the hypothesis space based on the given dialogue history. Again, we consider the Yes/No-annotation obtained from ChatGPT-Oracle as the ground truth and compute the hypothesis space at turn $HypSp_{t_i}$ by filtering out from $HypSp_{t_{i-1}}$ the items which do not have the property required at t_i. We compare the ground truth Hypothesis Space with a) the one generated by the external Guesser, ChatGPT-Guesser, and b) the one generated by the Questioner itself when prompted to explicitly update the list of candidates stepwise (ChatGPT-Q-stepwise). To this end, we compute the symmetric difference between the generated sets with the ground-truth ones for every question, and report the average symmetric difference of each game turn. The symmetric difference between two sets A and B is denoted by $A\Delta B$ and is defined as follows:

$$A\Delta B = (A - B) \cup (B - A)$$

For ChatGPT-Guesser, a high difference would mean that the model has difficulty integrating the information collected through the dialogue history. While for ChatGPT-Q-stepwise it would signal a difficulty in integrating the answer with the question turn by turn.

Search efficiency We measure the efficiency of the ChatGPT-Q's game strategy by computing the *average number of questions per game, (AQ)*. In addition to this, and as in (Ruggeri et al., 2016), we consider a question *unnecessary (UQ)*, if the preceding dialogue history already contained the information to identify the target. Again, we use the Y/N-annotation by ChatGPT-Oracle to determine whether this point has been reached by ChatGPT-Q.

The more the search strategy is effective, the shorter is the dialogue. The higher the number of UQ the closer is the model to pre-scholar children, who have been shown not to have learned the stopping rule yet. If the model asks just one UQ as last turn, that would still qualify it adult-like, since adults have been shown to ask a confirmation question before making the final guess (Testoni et al., 2022).

Experimental Settings We expect that bigger candidate sets could challenge the model's capacity to keep track of the information obtained through the dialogue, since they might require longer interactions. Moreover, with the GPT-based 8 cds the model should have all the knowledge to quickly arrive to identify the target. Hence, if ChatGPT's knowledge properly drives its question generation the dialogues of the games based on it should display an almost optimal information seeking strategy. Based on these conjectures, we compare the model when playing games whose hypothesis space a) consists of 8 and 16 candidate sets, b) is built out of McRae- or ChatGPT-feature norms. If the question generation is driven by the reasoning process on the space, we expect the model's

[6] A third type of questions are the *Pseudoconstraint-seeking (PCS)* which ask about a property but actually refer to only one item among the candidates. For the sake of simplicity, we do not consider them in our analysis, but see the Supplementary Material for statistics on them.

[7] We computed the EIG adapting the code by Testoni (2023).

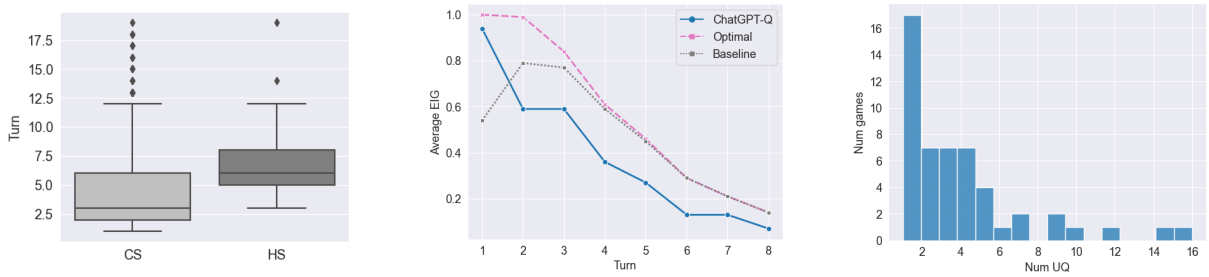


Figure 2: **Left**: ChatGPT-Q asks CS questions mostly in the earlier turns and HS in the later ones. Mann-Whitney U test shows the difference is statistically significant $p < 0.001$; **Middle**: ChatGPT's EIG is almost as high as the optimal model's EIG at the first turn, but it is lower than of the baseline's EIG at later turns. **Right**: Distribution of unnecessary questions.

performance to decrease when challenged with a higher number of candidates and to increase when the hierarchical structure of the candidates is based on the model's internal knowledge.

6 Results

In this section, we show the results we obtained following the experimental setup defined in Section 5. We start by evaluating ChatGPT's performance on the games with 8 candidates selected with the McRae-norms (8-McRae), and we then move to compare these results with those obtained by the model when challenged with spaces containing an higher number of candidates (16-McRae), or whose nested structure reflects the model's knowledge representation (8-GPT). Finally, we move to evaluate ChatGPT on the 8-McRae games when asked to play the game (ChatGPT-Q) and when asked to explicitly update the hypothesis space stepwise while playing the game (ChatGPT-Q-stepwise).

6.1 ChatGPT-Q on the McRae-8 games

Through the measures introduced above, here we aim to take a picture of how well and efficiently ChatGPT searches for information by considering McRae-8 games.

Information seeking strategy The results for the type of questions asked by ChatGPT-Q, the *optimal agent* and the *baseline* can be seen in Table 1. By construction, the optimal agent asks $log_2 N - 1$ CS question per game, followed by 1 or 2 HS questions (hence 1.5 on average), until it guesses the target; in other words, 57.14% of its questions are CS, and 42.86% are HS. Instead, the baseline asks only HS questions and, on average, it guesses the target in 4 questions. ChatGPT-Q asks mostly Constraint-Seeking questions (73.77%), it

8 cds based on McRae-norms			
	HS	CS	AQ
Optimal	42.86	57.14	3.50
Baseline	100	0	4.00
ChatGPT	26.33	73.77	7.24

Table 1: Information seeking strategy: Upper and lower bound of the overall percentage of hypothesis scanning (HS) vs. constraint seeking (CS) questions, and the average number of questions per game (AQ) -the difference is statistically significant based on the Wilcoxon signed-rank test (e.g., wrt the optimal agent, $p < 0.001$.

tends to ask CS questions in the early turns and HS questions towards the end of the dialogue – when indeed the latter becomes more efficient to split the space (Figure 2, left).

Moreover, our results show that the EIG of ChatGPT-Q's questions through the dialogue is far from the optimal agent's EIG (that half-split the space at each turn) and even lower than the baseline's (that splits the space into 1 vs. the other candidate at each turn) (see Figure 2, middle). Summing up, on the surface level, the strategy used by ChatGPT-Q reflects what an adult would do. However, the EIG analysis shows that ChatGPT-Q asks more uninformative questions compared both to the optimal agent and the baseline.

Hypothesis Space update We evaluate whether ChatGPT-Guesser is able to list the candidates that are still possible candidates based on the Question-Answer exchanges between ChatGPT-Q and ChatGPT-A. We do so by computing the difference, turn by turn, of such list with those considered as "ground truth" based on ChatGPT-Oracle annotations. The pattern we find (see the blue line in Figure 3, right) suggests the model has difficulty in integrating the information collected through the

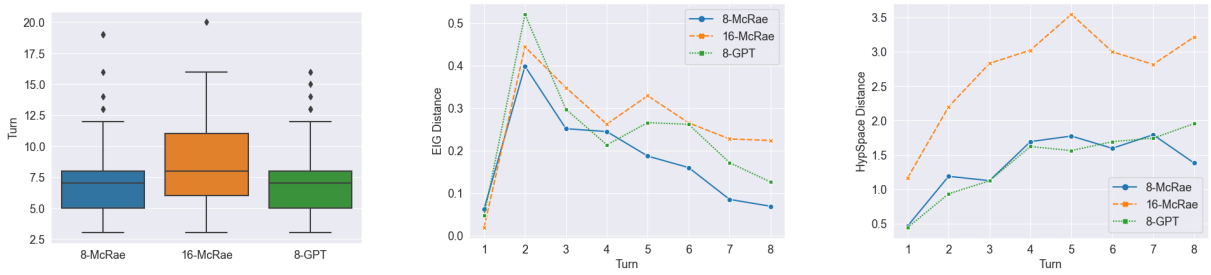


Figure 3: Increasing the candidates (**Blue vs. Orange**) causes the generation of longer dialogues (left) – maintaining a similar distance to the optimal model's AQ (8.67 vs 4.5); does not impact much the questions' EIG (middle), while it makes the guessing of the candidate by turn harder (right). Using candidates space structure based on GPT-norms (**Blue vs. Green**) does not impact any measure.

dialogue history; this weakness could impact its ability to stop asking questions when it has reached the singleton set.

Search efficiency By construction the optimal agent asks on average 3.5 questions per game, whereas the baseline asks on average 4 questions. ChatGPT-Q asks way more questions per game (7.24) compared to both models (Table 1). Such difference is statistically significant based on a Wilcoxon signed-rank test (e.g., wrt the optimal agent, $p < 0.001$). Moreover, ChatGPT-Q asks unnecessary questions (UQ) (questions asked after the singleton set has been reached) in 56.67% of games – 29.29% of its questions are unnecessary (See Figure 2, right for the distribution of UQ.) Summing up, in terms of search efficiency, ChatGPT's behavior is similar to that pre-scholar children who tend to not stop asking questions once there is only one item left in the hypothesis space.

6.2 Changing the games

Figure 3 illustrates how ChatGPT-Q's performance is effected by the Hypothesis space size (Blue vs. Orange) and of the features used to build it (Blue vs. Green). Hence, it compares the results discussed above (McRae-8) with those obtained in the other two settings: McRae-16 and GPT-8. In particular, it shows the comparison based on the number of questions per game (left), the distance between ChatGPT-Q and the optimal agent in terms of EIG (middle), and the average symmetric difference between the ground-truth hypothesis spaces update, based on ChatGPT-Oracle, and the one generated by ChatGPT-Guesser.

ChatGPT-Q on 16-McRae Given the difficulty the model has in keeping track of the space update, we expect that by increasing the number of candidates ChatGPT-Q' s performance will decrease. The results are not clear-cut: by moving from 8 to 16 candidates, the optimal agent would have an increase of 1 question per game, while ChatGPT increases of 1.43; the difference in terms of EIG is low, while ChatGPT-Guesser's performance deteriorates.[8]

ChatGPT-Q on 8-GPT The games built out of GPT-norms should reflect the model knowledge representation, therefore we expect that on the 8-GPT games ChatGPT-Q performance will increase. Instead, for none of the measures the difference is significant. This suggests that the feature norms used to build the hypothesis spaces do not impact the model's performance.

6.3 Changing the prompt

To further understand what causes ChatGPT inefficient strategy, we would need to run an ablation study by isolating the various processes that should be beyond the question generation. To simulate such study, we compare the set of results discussed so far on McRae-8 obtained by ChatGPT-Q simply prompted to play the game, with those obtained by ChatGPT- Q-stepwise, the model that is explicitly told to update the space turn by turn. Our results show that ChatGPT-Q-stepwise gets closer to the optimal model: it asks fewer questions per game (Figure 4, left) compared to ChatGPT-Q (6.4 vs. 7.24), the questions' EIG is higher across all the turns (Figure 4, middle), and it is more precise when updating the hypothesis space (Figure 4, right). This finding confirms the conjecture that ChatGPT main weakness lies in its difficulty in *mentally* updating the hypothesis space.

[8] Games generated on 16-McRae are significantly longer than those generated on 8-McRae, based on a Mann-Whitney U test ($p < 0.001$).

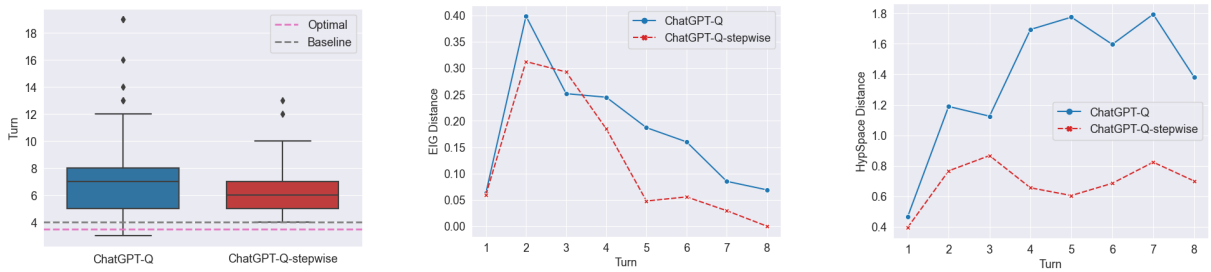


Figure 4: Changing the prompt improves the Questioner performance on all the measures, bringing it closer to the optimal model in terms of number of turns and EIG, an to the ground truth hypothesis space stepwise reduction.

ChatGPT-Q			
	CQ	TQ	SG
McRae-8	13.50	25.15	12.22
McRae-16	14.74	18.59	13.33
GPT-8	13.22	26.90	8.89
ChatGPT-Q-stepwise			
	CQ	TQ	SG
8-McRae	8.42	26.78	7.78

Table 2: Quality of the dialogue: Contradictory Questions (CQ) decrease when ChatGPT is asked to update the space explicitly. Yet, it is still unaware of Spoiled Games (SG).

7 Qualitative Analysis

In this section, we further dive into the quality of the dialogues generated by ChatGPT. First of all, we inspect whether it asks questions that do not reduce the space at all (trivial questions, TQ) or refer to candidates that have already been excluded in previous turns (contradictory questions, CQ). As we can see from the statistics reported in Table 2, ChatGPT is rather coherent through the dialogues; yet, the number of trivial questions is higher than what one would expect from a rational agent. In all the different settings, we observe a low peak in terms of EIG at the second turn. To understand the reason, we look into the 8-candidates sets (McRae and ChatGPT): 39.44% of the second questions are uninformative (EIG=0), with a large majority of trivial questions (92% of the uninformative). Interestingly, neither the size of the space nor the norms used to build it impact the number of trivial and contradictory questions. Instead, the coherence of the dialogues improves when the prompt is changed and ChatGPT is asked to update the space stepwise before asking the next question (ChatGPT-Q-stepwise on 8-McRae).

By inspecting the dialogues, we realized that in all the settings (8 vs 16 cds, McRae vs. ChatGPT-norms, explicit vs. implicit-update), there are games in which ChatGPT continues asking questions even when the Answerer has accidentally revealed the target (Spoiled games, SG – see an example in the Supplementary Material). This suggests that the model is pretending to play the game without having actually grasped the actual purpose of it. Most probably, a spoiler would not pass unobserved by a 4Y-child.

8 Conclusion

Our work shows that ChatGPT is able to identify superordinate features shared by items and ask questions that efficiently reduce the hypothesis space. At the first turn, it is close to an optimal agent using a half-split search. In later turns, however, it has difficulties making questions with respect to the updated space of the hypothesis. This weakness might be behind the high number of games in which it keeps on asking questions even though the dialogue history had led to identifying a possible target. We conjecture this behavior is not due to the lack of knowledge required by the game since it is displayed not only within the games based on McRae norms but also on those built out of GPT feature norms. Our conjecture is reinforced by the increased performance reached by the model when prompted to explicitly update the space before asking the next question. In this setting, the dialogue becomes shorter with fewer contradictory questions. Yet, even in such scenario, it does not notice when the Answerer reveals the target accidentally. Our results call for attention to modeling the human ability to keep a *mental scoreboard*, echoing what stated in Lewis (1979); Madureira and Schlangen (2022); Mazuecos et al. (2021). Finally, our work relates to the Chain-of-Thought (Wei et al., 2022) and similar prompting strategies, which we plan to investigate in the future.

9 Limitation

The backbone of the hierarchical space we built are feature norms. For the first level split, we used superordinates which by definition are disjoint. For the second level, we used all other feature norms by making sure that the feature that is shared by a subset is not listed in any of the members of the other subset; this process does not guarantee disjointedness of the two subsets. McRae features norms associated to a concept should be salient for it, while the absence of a feature from the list could be either because the feature does not hold for that concept or because it is not salient. Nevertheless we choose to use McRae-norms because they reflect human representation of the world and gave us the possibility of having a straight comparison with the games built out of ChatGPT-norms – comparison which shows that the knowledge used for building the hypothesis spaces does not impact the model's performance. We evaluated the model also on games built with taxonomic relations extracted from WordNet for both levels of the hierarchy: the patterns are very similar to those obtained with McRae- and GPT-norms (See the Supplementary Material for details.)

A second limitation is due to ChatGPT being a closed-source model, for which the exact training data is not known. We leave for future work the study of a LLM open source. Finally, we have not compared the model results on those that humans playing the games would achieve, instead we rely on the results obtained within the Cognitive Science literature about the 20-Q game.

Acknowledgements

We would like to thank Alberto Testoni for his contribution during the project design phase, and all the participants of the CLIC seminars at CIMeC for feedbacks during the early stage of the project. We are grateful to Dieu-Thu Le and the ARCIDUCA research group, led by Massimo Poesio, at Queen Mary University for their comments and suggestions on an early version of the paper. Finally, the last author is grateful to Amazon Alexa for the donation supporting research activities in her research group.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 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, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y.T. Lee, Y. Li, S. Lundberg, H. Nori, H. Planagi, M. T. Ribeiro, and Y. Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. ArXiv:2303.12712.

Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. In *2017 IEEE International Conference on Computer Vision*, pages 2951–2960.

H. Hansen and M. Hebart. 2022. Semantic features of object concepts generated with gpt-3. In *In Proceedings of the Annual Meeting of the Cognitive Science Society.*

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

D. Lewis. 1979. Scorekeeping in a language game. *Semantics from different points of view*, pages 172–187.

D. V. Lindley. 1956. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986 – 1005.

Brielen Madureira and David Schlangen. 2022. Can visual dialogue models do scorekeeping? Exploring how dialogue representations incrementally encode shared knowledge. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 651–664, Dublin, Ireland. Association for Computational Linguistics.

Mauricio Mazuecos, Franco M. Luque, Jorge Sánchez, Hernán Maina, Thomas Vadora, and Luciana Benotti. 2021. Region under Discussion for visual dialog. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4745–4759, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

K. McRae, G. S. Cree, M. S. Seidenberg, and C. McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 4(37):547–559.

B. Meder, J. D. Nelson, M. Jones, and A. Ruggeri. 2019. Stepwise versus globally optimal search in children and adults. *Cognition*.

E. Mosher and J. R Hornsby. 1966. On asking questions. *Studies in cognitive growth*.

S. Ott, K. Hebenstreit, V. Liévin, C. E. Hother, M. Moradi, M. Mayrhauser, R. Praas, O. Winther, and M. Samwald. 2023. Thoughtsource: A central hub for large language model reasoning data. Arxiv.2301.11596.

A. Rothe, B.M. Lake, and T.M. Gureckis. 2018. Do people ask good questions? *Comput Brain Behav*, 1:68–89.

A. Ruggeri, T. Lombrozo, T. L. Griffiths, and F Xu. 2016. Sources of developmental change in the efficiency of information search. *Developmental psychology*.

A. Ruggeri, C.M. Walker, T. Lombrozo, and A. Gopnik. 2021. How to help young children ask better questions? *Frontiers in Psychology*, 11. Doi: 10.3389/fpsyg.2020.586819.

Azzurra Ruggeri and Tania Lombrozo. 2015. Children adapt their questions to achieve efficient search. *Cognition*, 143:203–216.

A. Testoni. 2023. *Asking Strategic and Informative Questions in Visual Dialogue Games: Strengths and Weaknesses of Neural Generative Models*. Ph.D. thesis, IECS Doctoral School, University of Trento.

A. Testoni, C. Greco, and R. Bernardi. 2022. Artificial intelligence models do not ground negation, humans do. guesswhat?! dialogues as a case study. *Frontiers in big data*, 4(736709).

Peter Todd, Jean Ortega, Jennifer Davis, Gerd Gigerenzer, Daniel Goldstein, Adam Goodie, Ralph Hertwig, Ulrich Hoffrage, Kathryn Laskey, Laura Martignon, and Geoffrey Miller. 1999. *Simple Heuristics That Make Us Smart*. Oxford University Press.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

L. L. Wu and L. W. Barsalou. 2009. Perceptual simulation in conceptual combination: evidence from property generation. *Acta psychologica*.

D. Zhu, J. Chen, K. Haydarov, X. Shen, W. Zhang, and M. Elhoseiny. 2023. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. Arxiv.2303.06594.