

Supplementary Material

ChatGPT’s Information Seeking Strategy: Insights from the 20-Question Game

Leonardo Bertolazzi
University of Trento

Filippo Merlo
University of Trento

Davide Mazzaccara
University of Trento

Raffaella Bernardi
University of Trento

1 Hypothesis Space Construction

Table 1 reports the statistics about the distribution of feature per relation – for the GPT-3.5-turbo features, which are not superordinate, we do not have the corresponding WB relation.

The hypothesis space selection takes place through several levels in which starting from the total number of concepts, only those containing a randomly chosen first feature and not containing a random second feature are kept, grouped and contrasted with a second group of concepts that possess the second feature and do not possess the first. This procedure is repeated for each level of the hierarchy for each of the groups of concepts defined in the previous level. The features of the first level are always sampled from the list of features labeled as Superordinate features, while the features sampled in the lower levels are picked from all the other non Superordinate features. When a number of groups equal to half of the hypothetical space’s final size is reached, two concepts from each group are sampled without replacement and used to compose the hypothetical space. This sampling is done recursively, until all possible pairs have been extracted. These final pairs of concepts are not furtherly defined by some specific features since we assume that different concepts must have at least one feature differentiating them.

By following the procedure described above, for the 8 cds we obtained a total of 5184 sets and 600 sets for McRae-norms GPT-norms, respectively; for the 16 cds, we obtained 2384 sets for McRae-norms. For computational reasons, when dealing with the 16 cds (McRae norms) and the 8 cds based on GPT-norms we limited the recursion at the last level. Once all previous levels have been successfully passed. This limit amounted to 10 and 20 extractions for the 16 cds and the GPT-norms 8 cds, respectively.

Out of the 9 WB superordinates, 3 were not

Set	# Concepts	# Features
8-McRae	172	47
16-McRae	155	48
8-GPT	203	341

Table 2: Number of concepts and features used to create each game set.

frequent enough to contribute to the creation of the 16 candidate sets. Hence, to guarantee a fair comparison between the 8 and 16 cds, we reduced the former to the 90 sets built out of these 6 shared features. Table 1 lists the 6 features used.

Finally, the three different 90 sets utilized to create the games comprise 203 unique concepts and 341 unique features for the GPT-norms 8 cds, 172 unique concepts and 47 unique features for the McRae-norms 8 cds, and 155 unique concepts and 48 unique features for the McRae-norms 16 cds (see Table 2).

2 ChatGPT-A’s Error Rate

In section 4.1 we specified that a game ends when ChatGPT-A outputs “Yes, that’s correct”. However, sometimes ChatGPT-A outputs “Yes, that’s correct” even though ChatGPT-Q made a guess about an item which is not the target. Here is an example:

Target = socks

Answerer: This is the list of candidates: avocado, honeydew, bra, socks, cantaloupe, pear, cloak, coat.

Questioner: Is the item you were assigned a fruit?

Answerer: No, it is not a fruit.

Questioner: Is the item you were assigned a piece of clothing?

Answerer: Yes, it is a piece of clothing.

Questioner: Is the item you were assigned worn on the upper body?

WB relation	# Features	Names
superordinate	6	mammal, bird, clothing, weapon, fruit, vegetable
made_of	3	metal, wood, material
external_surface_property	18	small, large, long, round, different_colours, brown, black, white, green, loud, yellow, soft, red, hard, colourful, different_sizes, sharp, thin
external_component	8	4_legs, wings, legs, handle, a_tail, wheels, skin, eyes
entity_behavior	4	flies, eats, swims, lays_eggs
function	2	is_edible, used_for_killing
internal_surface_property	3	tastes_good, tastes_sweet, is_juicy
internal_component	1	has_seeds
location	2	lives_in_water, grows_in_gardens
GPT extra norms	335	contains_vitamins, eats_grass, is_source_of_fiber ...

Table 1: Details on the Wu and Barsalou’s relations and the feature norms used to build the game sets. The GPT norms do not have a corresponding WB relation, hence we consider them separately.

Answerer: No, it is not worn on the upper body.

Questioner: Is the item you were assigned worn on the lower body?

Answerer: No, it is not worn on the lower body either.

Questioner: Is the item you were assigned typically worn indoors?

Answerer: It can be worn both indoors and outdoors.

Questioner: Does the item you were assigned have sleeves?

Answerer: No, it does not have sleeves.

Questioner: Is the item you were assigned typically worn around the neck?

Answerer: No, it is not typically worn around the neck.

Questioner: Is the item you were assigned something that is worn to protect oneself from the cold or rain?

Answerer: Yes, it is often worn to protect oneself from the cold.

Questioner: Is the item you were assigned a type of jacket?

Answerer: No, it is not a type of jacket.

Questioner: Is the item you were assigned a shawl or a wrap?

Answerer: No, it is not a shawl or a wrap.

Questioner: Is the item you were assigned a hoodie?

Answerer: No, it is not a hoodie.

Questioner: Is the item you were assigned a vest?

Answerer: No, it is not a vest.

Questioner: Is the item you were assigned a scarf?

Answerer: Yes, that’s correct! The item assigned to me is socks.

Out of the dialogue generated with ChatGPT-Q and ChatGPT-A on the 8-McRae dialogues 12.22% are wrongly declared successful by ChatGPT-A.

3 Prompts

Considering the roles of system and user as in the ChatGPT API, the following section reports the prompts for ChatGPT-Q and ChatGPT-A.

ChatGPT-Q:

system: You are playing an interactive game with the user, who is assigned an item from a list of candidates. Ask as few questions as possible to identify the item, making only one question at each turn. The user can answer ‘yes’ or ‘no’.

user: This is the list of candidates: {candidates}.

ChatGPT-A:

system: You are playing an interactive game with the user, in which you are assigned one item from a list of candidates. The user will have to guess which one it is by asking yes/no questions, and you have to respond to each question only with ‘yes’ or ‘no’.

If the user correctly guesses your assigned item, respond with ‘Yes! That’s correct.’.

The item assigned to you is {target}.

ChatGPT-Oracle The Expected Information Gain is computed using the expected changes in the hypothesis space after each turn. Therefore, we need to know the answer to the question for each candidate in the set. To this end, we implement an External Oracle to collect this annotation, based on ChatGPT. ChatGPT-Oracle is instructed with the same ChatGPT-Q’s prompt (3); for each question in the dialogue, the External Oracle has to answer yes/no having one candidate at a time as target.

In different cases the ChatGPT-Oracle outputs a different answers from the expected yes/no (e.g., “I am sorry”). For all these cases, the Oracle is forced to output the expected answers appending a ‘Yes or No?’ at the end of the current question. If this strategy is not successful, the following shortened version of the prompt is provided:

You are playing an interactive game with the user, you have to answer each question about your item only with ‘yes’ or ‘no’.

ChatGPT-Guesser We use ChatGPT as a Guesser to spell out the referent set after each of the dialogue turns generated by ChatGPT-Q and ChatGPT-A. Differently from the Oracle, the External Guesser is given a sequence of Question Answer pairs, and the hypothesis space at turn 0. ChatGPT-Guesser prompt:

The user will provide you the history of a referential guessing game played by two agents called ‘assistant’ and ‘user’. Given the history of the conversation provided by the user, list the remaining candidates. You can only say ‘The remaining candidates are:’ followed by the list of remaining candidates.

This is the game history: <HISTORY>
List the remaining candidates.

ChatGPT-Q-stepwise To enhance ChatGPT’s logical reasoning, ChatGPT-Q-stepwise is fed with ChatGPT-Q’s prompt followed by a further instruction. The added instruction is as follows:

Format your output in the following manner: CANDIDATES: item, item, item, item ... QUESTION: text of the question

By incorporating this instruction, ChatGPT explicitly lists the candidates that have not been excluded from the game at each turn. This approach enables the model to constantly focus on an updated set of candidates without the need to “reason”

about the entire dialogue history when asking a new question.

4 Pseudoconstraint-seeking Questions

Regarding the *type of questions*, a more fine-grained analysis considers, along with *hypothesis scanning* and *constraint-seeking*, a third type of question, namely pseudoconstraint-seeking questions (PCS). PCS questions are similar to CS questions because they ask about a feature, but as HS questions refer to only one item among the candidates (e.g., *Does it has legs?*, with the hypothesis space of dog, a computer and a glass). PCS questions are frequent in adults, as an incorrect realization of a correct strategy. The following table shows the percentage of HS, CS, and PCS questions made by ChatGPT-Q using 8-McRae.

HS	CS	PCS
17.79	73.77	8.44

Table 3: PCS questions for the 8-McRae.

In Section 6, because the percentage of PCS questions was small (8.44%), we summed the percentages of HS and PCS questions, and for simplicity we reported them as HS. Instead, in comparing the position of HS and CS questions within a game, we excluded PCS from HS questions.

5 Spoiled Games

As mentioned in section 7, in all the settings in which we tested ChatGPT, there are games in which ChatGPT-Q continues asking questions even though the Answerer has accidentally revealed the target. The following is an example:

Target = blouse

Answerer: This is the list of candidates: dolphin, manatee, blouse, swimsuit, antelope, man, bikini, snowsuit.

Questioner: Is the item you’re thinking of an animal?

Answerer: No, it’s not an animal.

Questioner: Is the item you’re thinking of an article of clothing?

Answerer: Yes, it is an article of clothing.

Questioner: Is the item you’re thinking of specifically designed to be worn on the upper body?

Answerer: Yes, it is designed to be worn on the upper body.

Questioner: Is the item you're thinking of a type of blouse or shirt?

Answerer: Yes, the item I'm thinking of is a blouse.

Questioner: Great! Based on your answers, I think the item you are thinking of is a blouse. Is that correct?

Answerer: Yes! That's correct. Well done!

6 WordNet

The results of the main experiment show ChatGPT-Q is more informative in the first turn of the game, being able to recognise the best feature to divide the hypothesis space at this stage. This informativeness could depend on ChatGPT-Q's ability to better recognise superordinate features (i.e., taxonomic features, recall that only superordinate features are used to create the highest level of cds), or that no track of the space update is required at this stage of the game.

To observe the effect of superordinate vs non-superordinate features, we have employed WordNet to create 8 cds using only superordinate features. We annotated the same 541 concepts from McRae using WordNet hypernyms as features. For each concept, the first fourth inherited hypernyms are considered as features. For the concept *dog*, for example, the inherited hypernyms up to the fourth level are: *canine*, *carnivore*, *placental*, *mammal*. Among the features obtained, we selected a subset that were mutually exclusive: 6 for the first level and 3 hyponyms each for the second level of the hierarchy (see Tab. 4 for the features). This process results in 88 contrast sets with concepts presenting these features. We reached the 90 cds, as for McRae and ChatGPT norms, repeating the last two cds with a different random target.

I level	II level
vehicle	wheeled_vehicle, aircraft, vessel
plant	tree, herb, vegetable
bird	aquatic_bird, gallinaceous_bird, passerine
mammal	rodent, feline, canine
structure	housing, area, shelter
clothing	footwear, overgarment, undergarment

Table 4: selected features for Wordnet cds levels.

The results for WordNet cds are similar to the other 8 cds (McRae and GPT-norms), both in terms of average EIG (Fig. 1) and types of questions (Table 5). This makes us conclude ChatGPT-Q's

informativeness at the first turn does not mainly depend on its ability to recognise superordinate features.

8 cds based on WordNet		
	HS	CS
Optimal	42.86	57.14
Baseline	100	0
ChatGPT	25.75	74.25

Table 5: Information seeking strategy for WordNet cds compared to upper (Optimal) and lower bound (Baseline).

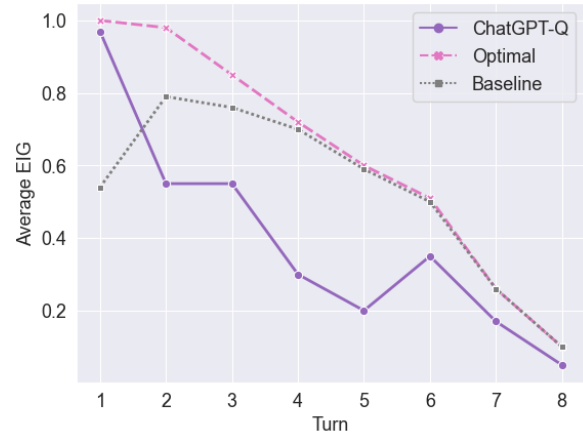


Figure 1: ChatGPT's EIG for the 8 cds with WordNet.