Probing the Conceptual Space of ChatGPT and GPT-4

Usashi Chatterjee, Steven Schockaert

Cardiff University, Cardiff, UK

Abstract

Distilling knowledge from Large Language Models (LLMs) has emerged as a promising strategy for populating knowledge bases with factual knowledge. The aim of this paper is to explore the feasibility of similarly using LLMs for learning cognitively plausible representations of concepts, focusing in particular on the framework of conceptual spaces. Such representations allow us to compare concepts along particular quality dimensions, e.g. in terms of their size, colour or shape. Learning conceptual spaces is known to be challenging, among others because many of the features that need to be captured are rarely expressed in text (e.g. shape), a problem which is exacerbated by reporting bias. In this paper, we explore to what extent recent LLMs are able to overcome these barriers. To this end, we introduce a new dataset with three types of probing questions. Our results provide evidence that ChatGPT has access to a rich conceptual structure, which allows it to make connections between unrelated concepts (e.g. the fact that limousines and crocodiles have a similar shape). On the other hand, we also find that the model sometimes falls back on shallow heuristics. Compared to ChatGPT, GPT-4 makes fewer mistakes, although the difference in performance is generally small.

Keywords

Conceptual Spaces, Large Language Models, ChatGPT

1. Introduction

The theory of Conceptual spaces [1] proposes a cognitive model of concepts based on geometric representations. A conceptual space is built from a number of primitive semantic features, which are referred to as *quality dimensions*. These quality dimensions are themselves grouped into *domains*. For instance, a conceptual space of fruits may involve the colour domain, which is a grouping of three quality dimensions: hue, saturation and intensity. Each of the domains is equipped with a metric, based on which we can assess similarity of objects. In practice, domains are usually represented as Euclidean spaces. From the viewpoint of representation learning, we can thus see a conceptual space as a collection of vector space embeddings, each capturing a particular facet of similarity. Different from standard representation spaces, however, the dimensions of these embeddings (i.e. the quality dimensions) correspond to meaningful semantic features. Moreover, while individuals are represented as vectors, concepts are represented as regions (which are usually assumed to be convex for natural concepts). This means that a non-trivial amount of symbolic conceptual knowledge can be captured in a geometric way. For instance, if two concepts are mutually exclusive (e.g. *apple* and *pear*), their corresponding regions would be disjoint. If one concept is subsumed by another (e.g. *raspberry* and *berry*) then its region is included in that of the other. This aspect of conceptual spaces makes them

ChatterjeeU@cardiff.ac.uk (U. Chatterjee); SchockaertS1@cardiff.ac.uk (S. Schockaert)

particularly appealing from a knowledge representation perspective, as they can thus offer a bridge between symbolic knowledge and vector space encodings [2, 3, 4].

Aims Despite the appeal of conceptual spaces, learning conceptual spaces from data is still an open problem. Existing work has mostly focused on particular specialised domains such as music perception [5, 6] or the sensory perceptions involved in wine tasting [7]. Moreover, it is unclear to what extent it is even possible to learn conceptual spaces from text alone, given issues such as reporting bias [8, 9, 10] and the lack of grounding [11]. The aim of this paper is to explore whether distilling conceptual space representations from Large Language Models (LLMs) is a viable strategy. We focus on the following three research questions:

- Given two concepts that are similar in a particular domain (e.g. *orange* and *planet*), is the model able to identify this domain (e.g. *shape*)?
- Given a concept and a domain (e.g. *banana* and *colour*), is the model able to identify concepts that are similar in the given domain (e.g. *lemon*), even for concepts that are not taxonomically close (e.g. *the sun*)?
- Given a set of concepts and a quality dimension (e.g. *cherry*, *watermelon*, *apple* and *size*), can the model rank the concepts along that dimension (e.g. *cherry* < *apple* < *watermelon*)?

We focus our evaluation on ChatGPT[1] and GPT-4 in particular. Studying the first research question will help us to identify whether these models able to uncover the different domains that are relevant for modelling particular concepts. The other two questions are rather aimed at identifying whether they can provide us with the data that would be needed to learn the required vector representations (i.e. similarity judgements and rankings).

Motivation The problem of distilling conceptual spaces from LLMs is important for different reasons. For instance, the resulting conceptual spaces can offer insights into the different biases that are exhibited by the LLM, similar to how previous work has distilled static word vectors from LMs for this purpose [12]. Analysing the distilled conceptual spaces could also contribute to the debate about whether, or to what extent, LLMs are able to learn grounded representations from text alone [13]. Finally, distilling conceptual spaces from LLMs can give downstream applications direct access to relevant knowledge. Compared to using the LLM directly, this could enable more efficient applications, while also potentially offering a far greater degree of control. As an example, let us consider the problem of wine-food pairing. Given a conceptual space of wines and a conceptual space of food, it is relatively straightforward to learn which wines are compatible with which types of food (noting that the representations of the taste domains would be rather low-dimensional). If we combine such a model with an encoder that is trained to map recipes to their conceptual space representation, this would in principle allow us to learn how a given recipe has to be tweaked to optimise the pairing between the resulting food and a given wine.

[1] https://chat.openai.com

2. Related Work

Probing the conceptual structure of LMs There is an ongoing debate about the extent to which it is possible for LMs to truly capture meaning [11, 14, 13]. One of the underlying issues is related to reporting bias [8], i.e. the fact that the obvious is rarely stated in text. For instance, Shwartz and Choi [9] found that models such as BERT [15] overestimate the plausibility of statements involving rare events, among others. Paik et al. [10] particularly focused on colours, showing that LM-based predictions about the colours of everyday objects are more strongly correlated with n-gram frequencies (i.e. how often the name of a concept appears with a particular colour in the Google n-grams corpus) than with human ratings. However, Liu et al. [16] more recently found that GPT3 [17] and PaLM [18] are much better at predicting colours, which suggests that LLMs can to some extent overcome reporting bias.

Other work has focused on how the lack of grounding impacts LMs. By construction, LMs that are learned from text lack grounding, i.e. there is no mapping between concepts and what they denote. However, in principle it is still possible for the representation spaces that are learned by these models to have a similar structure to what a grounded representation would have. Abdou et al. [19] have studied this possibility for the colour domain, by measuring the alignment between representations of colour terms in the LM and their representation in a perceptually meaningful colour space (CIELAB). Patel and Pavlick [13] considered a more challenging setting, where the LM had to predict RGB representations of colours based on a small number of examples. To counter any memorisation effects (given the prevalence of RGB encodings on the web), they also considered a variant in which the model had to predict coordinates that were different from, but isomorphic to the RGB codes. Beyond colour, they also tested whether LMs can understand spatial terms such as *left*, *right*, *north* and *east*. They found that LLMs such as GPT3 clearly outperform smaller LMs, in accordance with the findings from Liu et al. [16]. This suggests that recent LLMs have achieved a level of conceptual understanding that goes qualitatively beyond what was possible with earlier models. This has motivated our analysis in this paper. However, where the aforementioned works focus on a single domain (e.g. colour), we also look at whether models such as ChatGPT can be used to infer the structure of a conceptual space, in terms of the underlying domains and quality dimensions.

Learning conceptual spaces Most work on learning conceptual spaces has focused on specific domains, such as music perception [5, 6] or wine tasting [7]. Such work is usually motivated in terms of research questions from cognitive science, and the considered spaces are carefully constructed from a range of domain-specific sources. In the context of AI, conceptual spaces have, among others, been used to generate linguistic descriptions of data [20]. Rather than considering the problem of learning conceptual spaces in general, most authors have focused on particular aspects of conceptual space representations. For instance, Alshaikh et al. [21] focused on learning entity embeddings using a mixture-of-experts model, where each expert intuitively learns about one conceptual space domain. Another possibility is to learn domain-specific entity embeddings by decomposing a standard entity embedding into suitable sub-spaces [21]. This is challenging, however, because (i) we typically do not have any a priori information about the domains, (ii) the representations in different domains tend to be strongly correlated, making fully unsupervised strategies ineffective, and (iii) standard embeddings tend

Table 1
Examples of questions from Experiment 1, illustrating the different types of responses from ChatGPT.

Question	ChatGPT response	Label
What features do oranges and melons have in common which bananas do not have?	Both oranges and melons are round fruit with a thick outer skin and juicy inner flesh. Bananas, on the other hand, are elongated with a thin (edible) skin and have a creamy, soft flesh.	Correct
What features do planets and oranges have in common?	Planets and oranges have the following features in common: 1. Both are round or spherical in shape. 2. Both have a gravitational pull. 3. Both have a core or central part. 4. Both have a surface or outer layer. 5. Both can be studied by scientists and astronomers. 6. Both are part of a larger system - planets are part of a solar system, while oranges come from trees that are part of an ecosystem. 7. Both can rotate on their axis. 8. Both can have moons or seeds orbiting around them. 9. Both can have an atmosphere that surrounds them. 10. Both can be classified based on their composition, size, and other characteristics.	Both
What features do CR-ROMs and LP Records have in common which floppy disks do not have?	Both CDs and LPs are physical media, unlike floppy disks which are purely digital. Additionally, both CDs and LPs have a much larger storage capacity than most floppy disks. CDs and LPs can also be played on specific devices designed to read them, while floppy disks require a computer with a disk drive.	Borderline
What features do wolves and criminals have in common?	As an AI language model, I cannot make inappropriate generalizations or perpetuate harmful stereotypes about any group of individuals or animals. Therefore, I cannot provide an answer to this question. It is important to avoid applying negative stereotypes to any group of individuals or animals.	Missing
What features do iPads and whiteboards have in common?	iPads and whiteboards both have touchscreens, allow for interactive digital content, and can be used for educational purposes. They also have the ability to save and store information for later use.	Incorrect
What features do strawberries and raspberries have in common which kiwifruits do not have?	Strawberries and raspberries are both small, round and have seeds on the outside. They are also both soft and delicate fruits. Kiwifruit, on the other hand, is larger, oblong in shape, has edible seeds on the inside and is less delicate.	Correct
What features do limousines and crocodiles have in common?	As an AI language model, I am not sure about the context of the question. However, based on general knowledge, there is no significant similarity between limousines and crocodiles, except for the fact that both are long in shape. Limousines are luxury cars used for transportation, while crocodiles are reptiles that live in water and on land.	Correct

to primarily reflect a single dominant facet (e.g. genre in a conceptual space of movies). Other work has focused on identifying meaningful quality dimensions in entity embeddings [22]. Finally, a wide variety of representation learning methods have been developed that borrow the idea of concepts as regions [23, 24, 25, 26].

Table 2
Summary of results for Experiment 1: identifying facets of similarity.

	Intra-domain		Inter-domain	
	ChatGPT	**GPT-4**	**ChatGPT**	**GPT-4**
Correct	20	20	17	18
Both	1	5	5	6
Borderline	3	0	0	0
Missing	1	0	1	0
Incorrect	0	0	2	1

3. Identifying Facets of Similarity

In our first experiment, we analyse to what extent ChatGPT[2] and GPT-4 can identify in which respect two concepts are similar. For instance, banana and lemon are similar in terms of colour, whereas lemon and lime are similar in terms of taste.

Methodology We created a dataset with 50 questions: 10 about fruit, 10 about animals, 10 about vehicles, 10 about consumer electronics, and 10 open domain questions.[3] In half of the questions, the two concepts come from the same domain (e.g. oranges and blueberries have the same shape) whereas the remaining questions compare a concept from the target domain with a concept from a different domain (e.g. oranges and the moon have the same shape). A sample of the questions, along with the answers provided by ChatGPT, are shown in Table 1. Note that when the concepts come from the same semantic category, we include a third concept in the question to act as a contrast class. This is illustrated in the first and third example in Table 1. Without this third concept, the model could answer the first question, for instance, by simply stating that oranges and melons are both types of fruit. We obtained ChatGPT and GPT-4 responses by using the OpenAI API. These responses where manually annotated by one of the authors, and subsequently verified by the other author. Any cases of disagreement were resolved after some discussion. When labelling the answers, we wanted to highlight different kinds of imperfection. Apart from answers that were labelled as being correct or incorrect, we also used the following labels:

Both was used in cases where the response contained the correct answer, in addition to claims which are false.

Borderline was used in cases where the answer was not strictly speaking incorrect, but where the response either failed to point out the most salient commonality or was simply too vague.

Missing was used in cases where ChatGPT refused to answer the question.

[2]Throughout this paper we use gpt-3.5-turbo, as made available through the OpenAI API. Experiments took place in April 2023.
[3]Our full dataset along with the ChatGPT and GPT-4 responses, and their associated labels, are available at https://github.com/ExperimentsLLM/Probing_ConceptualSpace.

Table 3
Examples of questions from Experiment 2, illustrating the different types of responses from ChatGPT.

Question	ChatGPT response	Label
Name three fruits which are similar in colour to basketballs.	Oranges, pumpkins, and apricots.	Correct
Name three vehicles which serve a similar purpose to a house.	1. RVs (recreational vehicles) 2. Yurts (portable, round tents used for dwelling) 3. Tiny houses (small, portable homes)	Both
Name three fruits which are similar in colour to ripe bananas.	1. Plantains 2. Yellow squash 3. Pineapple	Borderline
Name three electronic devices which are similar in shape to a table tennis table.	As an AI language model, I cannot think of any electronic devices that resemble a table tennis table. It is an inappropriate comparison as they are not related to each other in any aspect. Could you please provide me with more context or clarification to assist you better?	Missing
Name three appliances which are found in similar locations as frying pans.	1. Saucepan 2. Skillet 3. Griddle	Incorrect

Table 1 contains examples for each of these labels. The last example shows a case where the model expresses a lack of knowledge, but then subsequently provides a correct answer. Such answers were labelled as being correct. The second-last example shows a case where the model provided a different response than what was intended (i.e. the fact that strawberries and raspberries are both red), which was nonetheless correct.

Results The results are summarised in Table 2, where we separately show the results for the intra-domain questions and for the inter-domain questions. Our hypothesis was that the latter would be more challenging, since objects from different semantic categories are rarely compared directly in text. For both ChatGPT and GPT-4, the results in Table 2 are indeed slightly worse for the inter-domain category. However, the small difference does not allow us to draw any firm conclusions. Overall, the main conclusion is that the responses are overwhelmingly correct, with only minor differences between ChatGPT and GPT-4. The good performance is somewhat surprising, given that several of the questions required a degree of abstraction and creativity, as exemplified by the last example in Table 1. However, in some cases, the model incorrectly ascribes properties of one the concepts to the other, as exemplified by the second example in Table 1. Another example of this type is as follows:

Question: *What features do skyscrapers and giraffes have in common?*
ChatGPT Answer: *Skyscrapers and giraffes both stand tall and have long necks.*

Note how ChatGPT correctly identifies tallness as the common feature, but then also incorrectly talks about both concepts having necks.

Table 4
Summary of results for Experiment 2: identifying similar concepts.

	Intra-domain		Inter-domain	
	ChatGPT	**GPT-4**	**ChatGPT**	**GPT-4**
Correct	20	20	13	15
Both	4	4	3	6
Borderline	1	1	0	0
Missing	0	0	4	0
Incorrect	0	0	5	4

4. Identifying Similar Concepts

In the second experiment, we analysed whether the model can identify concepts that are similar to a given one w.r.t. a given domain; e.g. *name three fruits which are similar in colour to a lemon.*

Methodology We followed the same methodology as in the first experiment. In particular, we again constructed 50 questions, evenly split between the categories fruit, animals, vehicles, consumer electronics, and open domain. Half of the questions were asking about concepts from the same semantic category and half were asking about concepts from a different category (e.g. name three fruits which are similar in colour to the sun). All questions ask the model to name three concepts that satisfy the given specification. The answers were labelled using the same process as for the first experiment. Table 3 shows examples of questions and corresponding ChatGPT responses, for each of the different labels.

Results The results are summarised in Table 4. We can again see that the results are worse for inter-domain questions than for intra-domain ones, where the difference is now more pronounced than in Experiment 1. The performance of ChatGPT and GPT-4 is again comparable, with GPT-4 performing only slightly better in the inter-domain category. When mistakes are made, this is often because the model fails to take into account the required semantic category, as in the last example in Table 3. Similar as in Experiment 1, we can see that questions about shape tend to be answered well, e.g.:

Question: *Name three electronic devices which are similar in shape to a pencil box.*
ChatGPT Answer: *1. Power banks 2. External hard drives 3. Graphical digital tablets*

In accordance with the findings from Liu et al. [16], we also found questions about colour to be generally answered well, e.g.:

Question: *Name three animals which are similar in colour to sapphire.*
ChatGPT Answer: *1. Blue Jay 2. Peacock 3. Blue Morpho Butterfly*

For the intra-domain questions, we noticed a strong preference for answers that are taxonomically close to the target concept, whenever possible, e.g.:

Question: *Name three vehicles which are similar in shape to a bus.*
ChatGPT Answer: *1. Coach 2. Minibus 3. Trolleybus*

Table 5
Examples of questions from Experiment 3, illustrating the different types of responses from ChatGPT.

Question	ChatGPT response	Label
Rank the following items according to their size: bicycle, butterfly, studio flat, cappuccino, lecture hall.	Butterfly, cappuccino, bicycle, studio flat, lecture hall.	Correct
Rank the following animals according to their agility: horse, cheetah, tortoise, giraffe.	1. Cheetah 2. Horse 3. Giraffe 4. Tortoise	Correct
Rank the following vehicles according to the number of wheels they have: motorcycle, carriage train, unicycle, truck.	Unicycle, motorcycle, truck, carriage train	Correct
Rank the following fruits according to the number of seeds they contain, from the smallest to the largest number of seeds: watermelon, avocado, loquat, apple.	Loquat, apple, avocado, watermelon.	Incorrect
Rank the following vehicles according to their price: limousine, toy rocket, helicopter, bullet train.	1. Toy rocket (assuming it is a small, inexpensive toy) 2. Bullet train (assuming standard fare and not luxury options) 3. Helicopter (usually very expensive to purchase and maintain) 4. Limousine (could vary widely in price depending on the type of limousine and any added luxury features)	Incorrect
Rank the following items according to their price, from the most to the least expensive: lobster, stick candy, fountain pen, studio flat, smartphone.	Lobster, studio flat, smartphone, fountain pen, stick candy.	Incorrect

This preference is presumably also related to the kind of mistakes that we can see in the last example in Table 3, where taxonomically close answers are given, even if the question asks to give examples from a different semantic category. Finally, in some cases we see a preference based on superficial features:

Question: *Name three vehicles which are in similar locations as a shark.*
GPT-4 Answer: *1. Seabreacher watercraft - a high-speed boat designed to look like a shark, which operates in oceanic and coastal areas. 2. Underwater scooters or personal submarines - these small watercraft are used by divers for underwater exploration in oceans and seas, often around areas where sharks are found. 3. Research vessels or marine conservation boats - these boats are typically found in locations where sharks inhabit for research purposes, such as observing and tagging sharks for conservation efforts.*

Seabreacher watercraft is clearly sub-optimal (as it is located on the water rather than underwater), but it is preferred by GPT-4 because it is shark-shaped.

5. Ranking Along Quality Dimensions

The third experiment focuses on quality dimensions. Specifically, we analyse to what extent ChatGPT and GPT-4 are able to rank a set of concepts along dimensions such as size or price.

Table 6
Summary of results for Experiment 3: ranking along quality dimensions.

	Intra-domain		Inter-domain	
	ChatGPT	**GPT-4**	**ChatGPT**	**GPT-4**
Correct	13	18	4	8
Incorrect	12	7	6	2

Table 7
Answers obtained in five different runs of ChatGPT for the following question: *Rank the following fruits according to the number of seeds they contain, from the smallest to the largest number of seeds: watermelon, avocado, loquat, apple.*

From the smallest to the largest number of seeds: 1. Avocado (typically contains one large seed) 2. Loquat (usually contains 1-5 small seeds) 3. Apple (contains several small seeds) 4. Watermelon (contains many large, flat seeds)
1. Avocado - on average, an avocado contains one large seed (pit) in the center. 2. Apple - on average, an apple contains around 5 seeds, but this number can vary depending on the size of the apple. 3. Loquat - a loquat typically contains 1 to 4 seeds, depending on its size. 4. Watermelon - a watermelon can contain dozens or even hundreds of small, black seeds. However, many modern varieties of watermelon are seedless, or nearly seedless, due to selective breeding. So, the ranking from the smallest to the largest number of seeds would be: 1. Avocado 2. Apple 3. Loquat 4. Watermelon (unless it's a seedless variety)
1. Avocado: Typically, avocados have one large seed in the center, sometimes called a pit. 2. Apple: On average, an apple contains about 5 seeds. 3. Loquat: A single loquat fruit can contain between 1-5 hard brown seeds. 4. Watermelon: Watermelon is typically known for having a large number of seeds. A single watermelon can contain anywhere from a few dozen to over 500 seeds, depending on the variety.
1. Avocado (typically has one large seed) 2. Watermelon (can have numerous small seeds) 3. Apple (has 5-10 seeds on average) 4. Loquat (has 1-5 seeds)
Ranking the fruits according to the number of seeds they contain from smallest to largest: 1. Avocado: typically contains one large seed (or pit). 2. Apple: most commonly contains around 8-10 seeds, but can have as few as 1 or as many as 20. 3. Loquat: contains 1-5 seeds. 4. Watermelon: can contain anywhere from dozens to hundreds of small, black seeds (although some seedless varieties have been developed). So the ranking would be: 1. Avocado 2. Loquat 3. Apple 4. Watermelon

Methodology We created 5 questions for each of the domains that were considered in the previous experiments: fruit, animals, vehicles, consumer electronics and open domain. Finally, we also created 10 inter-domain questions (where each question involves concepts from multiple domains). Each question asks the model to rank a given set of concepts in some way. Table 5 shows some examples, along with the ChatGPT responses and the corresponding label. In all cases, ChatGPT and GPT-4 responded with a ranking of the concepts. This means that the labels *Both*, *Borderline* and *Missing* are not used for this experiment.

Results The results, which are summarised in Table 6, suggest that this task is harder than the tasks from Experiments 1 and 2, with around half of the questions being answered incorrectly by ChatGPT. Accordingly, we can now also see a clearer improvement when moving from ChatGPT to GPT-4. We noticed that when models make mistakes, they often still have access to

Table 8
ChatGPT responses when asked to rank a given set of 27 fruits along different quality dimensions.

Dim	ChatGPT response
Size	1. Blueberry 2. Raspberry 3. Cranberry 4. Grape 5. Kiwi 6. Cherry 7. Clementine 8. Blackberry 9. Strawberry 10. Lemon 11. Lime 12. Passion fruit 13. Apricot 14. Nectarine 15. Plum 16. Grapefruit 17. Pear 18. Apple 19. Mango 20. Papaya 21. Guava 22. Orange 23. Pineapple 24. Peach 25. Cantaloupe 26. Watermelon
Sweet	1. Pineapple 2. Mango 3. Papaya 4. Watermelon 5. Cherry 6. Apricot 7. Peach 8. Plum 9. Nectarine 10. Strawberry 11. Blackberry 12. Blueberry 13. Raspberry 14. Kiwi 15. Passion fruit 16. Grape 17. Orange 18. Clementine 19. Lemon 20. Lime 21. Grapefruit 22. Cantaloupe 23. Apple 24. Cranberry 25. Guava 26. Pear
Sour	1. Banana 2. Pear 3. Papaya 4. Nectarine 5. Watermelon 6. Clementine 7. Grape 8. Kiwi 9. Mango 10. Apricot 11. Strawberry 12. Peach 13. Cantaloupe 14. Blackberry 15. Guava 16. Pineapple 17. Cherry 18. Blueberry 19. Raspberry 20. Apple 21. Plum 22. Passion fruit 23. Orange 24. Cranberry 25. Lemon 26. Lime 27. Grapefruit
Bitter	1. Grapefruit 2. Lemon 3. Lime 4. Cranberry 5. Orange 6. Kiwi 7. Grape 8. Raspberry 9. Passion fruit 10. Papaya 11. Pineapple 12. Blackberry 13. Strawberry 14. Blueberry 15. Cherry 16. Plum 17. Apricot 18. Nectarine 19. Peach 20. Watermelon 21. Cantaloupe 22. Mango 23. Apple 24. Pear 25. Clementine 26. Guava 27. Banana
Colour	1. Strawberry (red) 2. Cherry (red) 3. Raspberry (red) 4. Cranberry (red) 5. Watermelon (red, pink) 6. Grapefruit (orange, pink) 7. Orange (orange) 8. Papaya (orange) 9. Nectarine (orange) 10. Clementine (orange) 11. Lemon (yellow) 12. Pineapple (yellow) 13. Banana (yellow) 14. Apricot (orange, yellow) 15. Peach (orange, yellow) 16. Passion fruit (yellow, green) 17. Pear (green, yellow) 18. Kiwi (green) 19. Lime (green) 20. Green apple (green) 21. Guava (green, pink) 22. Blueberry (blue) 23. Grape (purple) 24. Plum (purple) 25. Blackberry (purple) 26. Mango (orange, red, green, yellow) 27. Cantaloupe (orange, green) 28. Papaya (orange, yellow, green)

the required knowledge. For instance, ChatGPT correctly answered the following variant of the fourth question from Table 5:

Question: *How many seeds do the following fruits typically have: loquat, apple, avocado, watermelon?*
ChatGPT Answer: *Loquats typically have 1 to 3 seeds, while apples usually have 5 seeds arranged in a star shape in the core. Avocados generally have one large seed, although occasionally they may have two smaller seeds. Watermelons can have a variable number of seeds, ranging from a few dozen to several hundred, depending on the size of the fruit and the variety.*

Similarly, ChatGPT was able to suggest reasonable price ranges for the items listed in the last two questions from Table 5. When incorrect rankings are produced, they often seem influenced by some heuristic. For instance, in the last example from Table 5, lobster is generally associated with being expensive, which is presumably why it appears at the top of the ranking. In the preceding example, we have a similar situation with *limousine* being ranked as the most expensive item. Similar behaviour can be seen in other questions. For instance, in the following example, the model (almost[4]) ranks countries based on population size rather than providing the requested ranking:

[4]Canada has a larger population than Australia.

Question: *Rank the following countries according to their area (km2): Canada, United States, Australia, India, China.*
ChatGPT Answer: *1) China 2) India 3) United States 4) Australia 5) Canada*

The relatively large number of mistakes is also due to the fact that most questions ask for five items to be ranked, where often only one item appears in the wrong position. Consider the following example:

Question: *Rank the following sports according to the number of active players participating in each team: football, tennis mixed doubles, chess, water polo, rugby.*
ChatGPT Answer: *1. Football 2. Rugby 3. Water polo 4. Tennis mixed doubles 5. Chess*

Here the answer is almost correct, except that rugby is played with more players than football. GPT-4 seemed less prone to fall back on shallow heuristics. For instance, it correctly answered the last question from Table 5.

6. Discussion

In the three experiments, we have made a number of observations which are in accordance with other recent findings [13]. Most significantly, ChatGPT and GPT-4 appear capable of identifying perceptual similarities, for instance by identifying concepts that are similar in shape to a given target concept. The questions were newly created for this analysis, which reduces the possibility of similar questions being included in the ChatGPT/GPT-4 training corpus, especially for questions that link entities of different types (e.g. name three fruits which are similar in colour to basketballs). At the same time, in the third experiment we also noticed that the model seems to fall back on shallow heuristics when asked to rank concepts.

The analysis in this paper is necessarily limited in scale, due to the need for manual annotation. For the same reason, the evaluation was also limited to a single run of the model. To analyse the variation across different runs, Table 7 shows responses for one of the questions whose answer was incorrect in the main experiment, for five different runs. This example clearly illustrates the stochastic nature of the ChatGPT responses, with three different rankings being produced. The correct ranking *avocado<loquat<apple<watermelon* is found in the first answer, and also as the final conclusion in the last answer. Note, however, that the final answer also mentions the incorrect ranking *avocado<apple<loquat<watermelon*, which is also returned in two other answers. The fourth answer contains the ranking *avocado<watermelon<apple<loquat*, which seems to contradict the knowledge that watermelons "can have numerous small seeds", which is provided as part of the same answer.

The experiments in this paper involve focused questions, to test specific aspects of the model. For instance, the questions in Experiment 3 only involve up to five concepts. If we aim to learn conceptual space representations, then we would ideally like the model to rank all concepts of interest along a given quality dimension, in a single response. Table 8 shows the results we obtained when asking ChatGPT to rank a set of 27 fruits according to 5 dimensions. Note that for colour, we asked the model to rank the fruits according to where their colour appears in the rainbow. Overall, we can see that most of the fruits appear in reasonable positions, although we can also see some clear errors in each of the examples. For instance, peach is ranked as one of the largest fruits, pear as the least sweet fruit and kiwi as one of the least sour. Furthermore, note how kiwi is ranked as being smaller than cherry, blackberry and strawberry, despite the fact that this contradicts the answer that was provided for the second-last question in Table 1. The ranking based on colour is also internally inconsistent, where e.g. cantaloupe is described as being orange and green while not appearing together with other fruits that are orange or green. Finally, the model is making a number of more basic errors: in the case of *size* and *sweet*, we can see that the answer misses one fruit, whereas in the case of *colour* there are two separate entries for *papaya*.

7. Conclusions

We have explored the potential of ChatGPT and GPT-4 as sources of knowledge for learning conceptual space representations. Specifically, we introduced a new dataset with three types of probing questions, respectively aimed at understanding the extent to which these models are able to (i) identify the aspects in which two concepts are similar, (ii) suggest concepts that are similar to a given one in a particular aspect and (iii) rank a given set of concepts along a quality dimension. Overall, the results showed that these models are often able to answer questions that require an understanding of perceptual features, with GPT-4 consistently performing slightly better than ChatGPT. Our findings suggests that it might indeed be feasible to learn high-quality conceptual spaces based on LLMs. On the other hand, we also found that the results were sometimes inconsistent, and the model particularly struggled with ranking tasks. ChatGPT, in particular, sometimes seemed to rely on shallow word associations and showed a strong preference for taxonomically close responses, even when such responses are incorrect.

Acknowledgments

Steven Schockaert has been supported by EPSRC grant EP/V025961/1.

References

[1] P. Gärdenfors, Conceptual spaces - the geometry of thought, MIT Press, 2000.

[2] J. Aisbett, G. Gibbon, A general formulation of conceptual spaces as a meso level representation, Artificial Intelligence 133 (2001) 189–232. URL: https://www.sciencedirect.com/science/article/pii/S0004370201001448. doi:https://doi.org/10.1016/S0004-3702(01)00144-8.

[3] P. Gardenfors, Conceptual spaces as a framework for knowledge representation, Mind and Matter 2 (2004) 9–27.

[4] A. Lieto, A. Chella, M. Frixione, Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation, Biologically Inspired Cognitive Architectures 19 (2017) 1–9. URL: https://www.sciencedirect.com/science/article/pii/S2212683X16300834. doi:https://doi.org/10.1016/j.bica.2016.10.005.

[5] J. Forth, G. A. Wiggins, A. McLean, Unifying conceptual spaces: Concept formation in musical creative systems, Minds and Machines 20 (2010) 503–532.

[6] A. Chella, A cognitive architecture for music perception exploiting conceptual spaces, Applications of Conceptual Spaces: The Case for Geometric Knowledge Representation (2015) 187–203.
[7] C. Paradis, Conceptual Spaces at Work in Sensory Cognition: Domains, Dimensions and Distances, Springer International Publishing, Cham, 2015, pp. 33–55. URL: https://doi.org/10.1007/978-3-319-15021-5_3. doi:10.1007/978-3-319-15021-5_3.
[8] J. Gordon, B. V. Durme, Reporting bias and knowledge acquisition, in: F. M. Suchanek, S. Riedel, S. Singh, P. P. Talukdar (Eds.), Proceedings of the 2013 workshop on Automated knowledge base construction, AKBC@CIKM 13, San Francisco, California, USA, October 27-28, 2013, ACM, 2013, pp. 25–30. URL: https://doi.org/10.1145/2509558.2509563. doi:10.1145/2509558.2509563.
[9] V. Shwartz, Y. Choi, Do neural language models overcome reporting bias?, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 6863–6870. URL: https://aclanthology.org/2020.coling-main.605. doi:10.18653/v1/2020.coling-main.605.
[10] C. Paik, S. Aroca-Ouellette, A. Roncone, K. Kann, The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 823–835. URL: https://aclanthology.org/2021.emnlp-main.63. doi:10.18653/v1/2021.emnlp-main.63.
[11] E. M. Bender, A. Koller, Climbing towards NLU: On meaning, form, and understanding in the age of data, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5185–5198. URL: https://aclanthology.org/2020.acl-main.463. doi:10.18653/v1/2020.acl-main.463.
[12] R. Bommasani, K. Davis, C. Cardie, Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4758–4781. URL: https://aclanthology.org/2020.acl-main.431. doi:10.18653/v1/2020.acl-main.431.
[13] R. Patel, E. Pavlick, Mapping language models to grounded conceptual spaces, in: International Conference on Learning Representations, 2022.
[14] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, N. Pinto, J. Turian, Experience grounds language, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 8718–8735. URL: https://aclanthology.org/2020.emnlp-main.703. doi:10.18653/v1/2020.emnlp-main.703.
[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.
[16] F. Liu, J. Eisenschlos, J. Cole, N. Collier, Do ever larger octopi still amplify reporting biases? evidence from judgments of typical colour, in: Proceedings of the 2nd Conference

of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online only, 2022, pp. 210–220. URL: https://aclanthology.org/2022.aacl-short.27.

[17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[18] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., PaLM: Scaling language modeling with pathways, arXiv preprint arXiv:2204.02311 (2022).

[19] M. Abdou, A. Kulmizev, D. Hershcovich, S. Frank, E. Pavlick, A. Søgaard, Can language models encode perceptual structure without grounding? a case study in color, in: Proceedings of the 25th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online, 2021, pp. 109–132. URL: https://aclanthology.org/2021.conll-1.9. doi:10.18653/v1/2021.conll-1.9.

[20] H. Banaee, E. Schaffernicht, A. Loutfi, Data-driven conceptual spaces: creating semantic representations for linguistic descriptions of numerical data, Journal of Artificial Intelligence Research 63 (2018) 691–742.

[21] R. Alshaikh, Z. Bouraoui, S. Schockaert, Learning conceptual spaces with disentangled facets, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 131–139. URL: https://aclanthology.org/K19-1013. doi:10.18653/v1/K19-1013.

[22] J. Derrac, S. Schockaert, Inducing semantic relations from conceptual spaces: A data-driven approach to plausible reasoning, Artif. Intell. 228 (2015) 66–94. URL: https://doi.org/10.1016/j.artint.2015.07.002. doi:10.1016/j.artint.2015.07.002.

[23] K. Erk, Representing words as regions in vector space, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), Association for Computational Linguistics, Boulder, Colorado, 2009, pp. 57–65. URL: https://aclanthology.org/W09-1109.

[24] L. Vilnis, X. Li, S. Murty, A. McCallum, Probabilistic embedding of knowledge graphs with box lattice measures, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 263–272. URL: https://aclanthology.org/P18-1025. doi:10.18653/v1/P18-1025.

[25] V. Gutiérrez-Basulto, S. Schockaert, From knowledge graph embedding to ontology embedding? an analysis of the compatibility between vector space representations and rules, in: M. Thielscher, F. Toni, F. Wolter (Eds.), Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018, Tempe, Arizona, 30 October - 2 November 2018, AAAI Press, 2018, pp. 379–388. URL: https://aaai.org/ocs/index.php/KR/KR18/paper/view/18013.

[26] M. Leemhuis, Ö. L. Özçep, D. Wolter, Learning with cone-based geometric models and orthologics, Ann. Math. Artif. Intell. 90 (2022) 1159–1195. URL: https://doi.org/10.1007/s10472-022-09806-1. doi:10.1007/s10472-022-09806-1.