Human-like property induction is a challenge for large language models

Simon Jerome Han (simon.jerome.han@gmail.com)
Keith J. Ransom (keith.ransom@unimelb.edu.au)
Andrew Perfors (andrew.perfors@unimelb.edu.au)
Charles Kemp (c.kemp@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne, Parkville, Victoria 3010, Australia

Abstract

The impressive recent performance of large language models such as GPT-3 has led many to wonder to what extent they can serve as models of general intelligence or are similar to human cognition. We address this issue by applying GPT-3 to a classic problem in human inductive reasoning known as property induction. Our results suggest that while GPT-3 can qualitatively mimic human performance for some inductive phenomena (especially those that depend primarily on similarity relationships), it reasons in a qualitatively distinct way on phenomena that require more theoretical understanding. We propose that this emerges due to the reasoning abilities of GPT-3 rather than its underlying representations, and suggest that increasing its scale is unlikely to change this pattern. **Keywords:** reasoning; property induction; neural networks; GPT-3; AI

Introduction

In recent years, transformer-based language models (TLMs) have attracted interest for their impressive performance on a wide range of language tasks including translation, summarisation and question answering. Language models such as GPT-3 (Brown et al., 2020) and Gopher (Rae et al., 2021) are so adept at engaging in apparently natural conversations on a broad range of topics that it is tempting to conclude that they show some degree of general intelligence, and thus that they are potentially useful as models of human cognition.

This possibility has given rise to an active research area aiming to probe the scope and limitations of the current generation of TLMs, as well as to anticipate the abilities of future generations that are even more powerful. Many families of tasks are used in this literature, including some that specifically target linguistic abilities (Hu, Gauthier, Qian, Wilcox, & Levy, 2020) and others that target commonsense knowledge and logical reasoning (Rae et al., 2021). Here we propose that the set of existing tasks can be usefully supplemented by drawing on the extensive psychological literature on inductive reasoning. To support this general claim we explore the extent to which one prominent TLM (GPT-3) is able to account for core phenomena in human property induction.

Inductive reasoning is one of the most central cognitive tasks people face. It involves arriving at plausible conclusions in the face of uncertainty, and is typically involved when dealing with sparse or noisy data. In a property induction task (Rips, 1975), people are given premises that indicate that a property is shared by one or more categories (e.g. MICE and SQUIRRELS have sesamoid bones) and must assess whether the property is shared by a different category (do POSSUMS have sesamoid bones?). The task is simple and has been used to study the reasoning of children (Carey, 1985) and adults from a broad range of cultural backgrounds (López, Atran, Coley, Medin, & Smith, 1997). Despite this apparent simplicity, the task yields a rich range of phenomena that draw on many kinds of knowledge (for a review, see Hayes and Heit, 2018). This knowledge includes not just similarity (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990), but also causal relationships (Medin, Coley, Storms, & Hayes, 2003) and assumptions about the process by which the premises were generated (Ransom, Perfors, & Navarro, 2016).

The range of inductive phenomena – from simple similarity-based effects to theory-based effects that draw on richer kinds of knowledge – corresponds to a sequence of increasingly difficult challenges for TLMs and other computational approaches (Sloman, 1993; Rogers & McClelland, 2004; Kemp & Tenenbaum, 2009). As such, property induction tasks could potentially lead to benchmarks that help to drive continued progress in computer science and AI. Indeed, some of the benchmarks currently used to evaluate TLMs focus on inductive problems (Sap, Rashkin, Chen, LeBras, & Choi, 2019). However, as far as we know, property induction has not been considered at all when evaluating TLMs.

For psychologists, property induction is relevant to a literature that assesses TLMs and predecessors such as LSA (Landauer & Dumais, 1997) as computational accounts of the acquisition, use, and representation of semantic knowledge. Recent work has evaluated the extent to which TLMs account for human similarity ratings, typicality ratings, and response times (Bhatia & Richie, 2021; Lake & Murphy, 2021), but there has been relatively little work on inductive reasoning. A notable exception is the work of Misra, Ettinger, and Taylor Rayz (2021), who focus on typicality and include property induction as one of the tasks that they consider. Typicality is among the phenomena considered here, but we investigate many others as well.

The next section introduces the inductive phenomena that we analyse, along with a theoretical account of these phenomena known as the Similarity Coverage Model (SCM). We then compare the inferences of GPT-3 with humans on these phenomena (Osherson et al., 1990). We find that GPT-3 accounts for some aspects of human inductive reasoning, but overall the match between GPT-3 and humans is relatively poor. Our results suggest that the primary shortcomings of GPT-3 lie in the inferential processes it carries out over its representations rather than the representations themselves. Our final analysis suggests that simply increasing the scale of GPT-3 is unlikely to allow it to attain human-level inductive abilities, and we conclude by discussing implications and identifying directions for future work.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Table 1: Eleven property induction phenomena introduced by Osherson et al. (1990) and investigated in this paper. The second column is based on the levels occupied by premises and conclusion in a category hierarchy. For specific arguments, premises and conclusion lie at the same level, but for general arguments the conclusion lies at a higher level than the premises.

Phenomenon	Type	Stronger argument	Weaker argument
Premise-Conclusion Similarity	Specific	{ROBIN, BLUEJAY} → SPARROW	{ROBIN, BLUEJAY} → GOOSE
Premise Typicality	General	ROBIN → BIRD	PENGUIN → BIRD
Conclusion Specificity	General	{BLUEJAY, FALCON} → BIRD	{BLUEJAY, FALCON} → ANIMAL
Premise Monotonicity	General	{SPARROW, EAGLE, HAWK} → BIRD	{SPARROW, EAGLE} → BIRD
Premise Monotonicity	Specific	{PIG, WOLF, FOX} → GORILLA	{PIG, WOLF} → GORILLA
Premise Diversity	General	{HIPPO, HAMSTER} → MAMMAL	{HIPPO, RHINO} → MAMMAL
Premise Diversity	Specific	{LION, GIRAFFE} → RABBIT	{LION, TIGER} → RABBIT
Non-Monotonicity	General	{CROW, PEACOCK} → BIRD	{CROW, PEACOCK, RABBIT} → BIRD
Non-Monotonicity	Specific	FLY → BEE	{FLY, ORANGUTAN} → BEE
Premise-Conclusion Asymmetry	Specific	MICE → BAT	BAT → MICE
Inclusion Fallacy	Both	ROBIN → BIRD	ROBIN → OSTRICH

Inductive Phenomena

We follow a long tradition of studies that examine inductive reasoning by focusing on property induction with semantically "blank" or unfamiliar properties. In a typical property induction task, participants are asked to rate the strength of inductive arguments like "ROBINS have property P, therefore BIRDS have property P." We will use the notation ROBIN → BIRD to indicate that an argument involves generalizing a property from ROBIN to BIRDS in general.

Although this task may seem simple, it gives rise to numerous phenomena that are indicative of the ways in which humans reason inductively. Osherson et al. (1990) present thirteen such phenomena, and eleven of the thirteen are shown in Table 1. All eleven are illustrated by comparing a stronger argument with a weaker argument, and the two phenomena not included in the table are omitted because they are not formulated in terms of a similar comparison.

Some of the phenomena directly capture effects of similarity or typicality. For instance, **Premise-Conclusion Similarity** reflects the finding that people are more likely to generalise a property from one concept to another when the concepts are more similar. **Premise Typicality** is the finding that arguments are stronger if the premises are more typical of the conclusions. A slightly less reliable phenomenon, **Premise-Conclusion Asymmetry**, reflects the fact that an argument that generalises from a typical category member to a less typical one (e.g. MICE → BATS) is often rated as stronger than the reverse argument (e.g. BATS → MICE) because atypical categories are more likely to have atypical properties.

Other phenomena relate to the hierarchical organization of categories. **Conclusion Specificity** reflects the intuition that greater inductive leaps are required to support broader generalisations; arguments are thus stronger if the conclusion category is more specific. The **Inclusion Fallacy** relates to the observation that a general argument from a category to its enclosing class (e.g. ROBIN → BIRD) can appear stronger than a more specific argument (e.g. ROBIN → OSTRICH) that is nonetheless logically entailed by the first. The inclusion fallacy appears in Table 1 for completeness, but because it is normally viewed as a fallacy it may not necessarily be appropriate as a target for AI models like GPT-3.

There are also phenomena which appear to reflect more sophisticated or theory-based reasoning about underlying mechanisms. **Premise Diversity** refers to the fact that arguments are more compelling if their premises are less similar to one another. This captures the general intuition, based on an understanding of statistical sampling, that diverse evidence is more compelling than narrow evidence. A similar mechanism may underlie systematic violations of **Premise Monotonicity**, which is the phenomenon that additional positive premises increase the strength of an argument. Premise Monotonicity often holds if all premises are drawn from the same superordinate category, but adding premises from a different superordinate category can lead to **Premise Non-Monotonicity**. For example, the inclusion of ORANGUTAN in the argument {FLY, ORANGUTAN} → BEE means that the context of the argument (the smallest category which includes the premise and inclusion categories) changes from INSECT to ANIMAL. This suggests that the property in question is not insect-specific, and thus reduces the chance that bees share it. These systematic violations of premise monotonicity and premise diversity have been shown to be influenced by the reasoner's theoretical assumptions about how the premises were generated (Ransom et al., 2016; Hayes, Navarro, Stephens, Ransom, & Dilevski, 2019).

Similarity-Coverage Model

In addition to characterizing the inductive phenomena just described, Osherson et al. (1990) presented a theory known as the Similarity Coverage Model (SCM) that is able to account for all of these phenomena. We introduce the SCM here because it will be used as part of our evaluation of GPT-3.

The SCM builds on the fact that several inductive phenomena can be derived purely from concept similarity. For example, ROBIN → SPARROW is stronger than ROBIN → GOOSE because robins are more similar to sparrows than geese. Sim-

ilarly, ROBIN → BIRD is stronger than PENGUIN → BIRD because robins are more similar to the prototypical bird than penguins are. In both cases, the probability that the premise and conclusion categories share a property increases solely based on the similarity of the two sets of categories.

Although similarity based accounts of property induction are simple and intuitive, they fail to account for more complex phenomena such as non-monotonicity and diversity. The SCM accounts for these phenomena by incorporating a notion called *coverage*, which denotes the degree to which the premise categories are similar to members of the lowest level category class that encapsulates each of the premise and conclusion categories. Osherson et al. (1990) demonstrate that a weighted combination of coverage and premise-conclusion similarity captures all eleven of the phenomena in Table 1.

Comparing GPT-3 with humans

In order to assess the extent to which GPT-3 captures people's judgments, we need a principled way to elicit its responses.

Presenting arguments to GPT-3

Because effective prompt design is a critically important aspect of interacting with GPT-3, we experimented with multiple prompts. This included a question-answer format, a conditional format akin to that used by Misra et al. (2021), and a format that omitted properties and simply listed a sequence of premise categories followed by a conclusion. We also experimented with including written task instructions within the prompt and varied whether we asked GPT-3 for direct completions or instead provided it with a predetermined set of answers. Different prompts led to slightly different patterns of responses, but our general conclusions about the limitations and abilities of GPT-3 are broadly consistent no matter what prompts were used. As a result, we report the single prompt design that elicited the most human-like performance.

The best-performing prompt design was a question-answer based prompt that included a task description and contextual information followed by a yes/no entailment question. We used a feature of the GPT-3 API[1] that allowed us to extract the probability assigned by the model to a particular word after it had seen some preceding context. For example, to obtain a strength rating for the argument DOGS → BEARS, the model was given the text:

```
You are an expert on the properties
that animals have, and you understand
how animals share properties in common.
Recently some animals have been discovered
to have property P. We know that dogs have
property P. Does this mean that bears have
property P? Please answer 'Yes' or 'No'.
```

[1]Interaction with GPT-3 was carried out via the Python "OpenAI" library using the `text-davinci-001` engine, the most capable GPT-3 model available at the time. To eliminate stochasticity, we set temperature $t = 0$.

The final token in the answer was then either `Yes` or `No`, and the probability assigned to `Yes` relative to `No` was taken as its rating of the strength of the argument.

Does GPT-3 account for individual phenomena?

Using this approach, we now ask whether GPT-3 is sensitive to the eleven individual phenomena in Table 1. For each, Osherson et al. (1990) presented participants with the pairs of arguments in the table and asked them to choose the argument whose premises "provide a better reason for believing its conclusion." The proportions of people who preferred the stronger argument are shown as black dots in Figure 1. For example, for Premise-Conclusion Similarity, around 90% of people indicated that they thought {ROBIN, BLUEJAY} → SPARROW was stronger than {ROBIN, BLUEJAY} → GOOSE.

As an initial test, we gave the same pairs of arguments to GPT-3 and asked it to choose the stronger of each pair. To allow for a comparison between GPT-3 and human responses, for each argument pair we took the strength rating that GPT-3 assigned to the stronger argument and divided it by the sum of strength ratings assigned by GPT-3 to both arguments; this corresponds to the white dots in Figure 1. All responses are relatively close to 0.5, but this could simply reflect different scaling. The more interesting question is thus whether the model response exceeds 0.5 (i.e., indicating that the model prefers the same argument that people think is stronger). Based on the white dots, it appears that while GPT-3 may capture a few of the phenomena, it struggles on most of them. Each dot is based on a single argument pair, however, and we are wary of drawing strong conclusions about any particular phenomenon on that basis.

We therefore performed a more systematic test by generating a larger set of argument pairs for each phenomenon. These arguments involved the 129 animals included in the Leuven Natural Concept database (De Deyne et al., 2008), which are grouped into five superordinate categories (MAMMALS, BIRDS, FISH, INSECTS, and REPTILES). For each phenomenon we generated 100 argument pairs that followed the same basic template as shown in Table 1. For example, each pair for Premise-Conclusion Similarity included two arguments with matching premises and within each of these pairs the premises and conclusion were drawn from the same superordinate category.

In order to evaluate GPT-3 on these argument pairs it was necessary to determine which member of each pair was stronger. We therefore followed Osherson et al. (1990) and classified arguments as stronger or weaker on the basis of the predictions of the SCM. For most phenomena we were able to directly calculate SCM scores for both arguments in a pair using pairwise similarity ratings obtained from the same database the animals were sampled from (De Deyne et al., 2008). However, this dataset only includes ratings between pairs of categories within the same superordinate class, which meant that SCM scores could not be obtained for both members of the argument pairs for Conclusion Specificity and Non-Monotonicity. In these cases, however, it is straight-

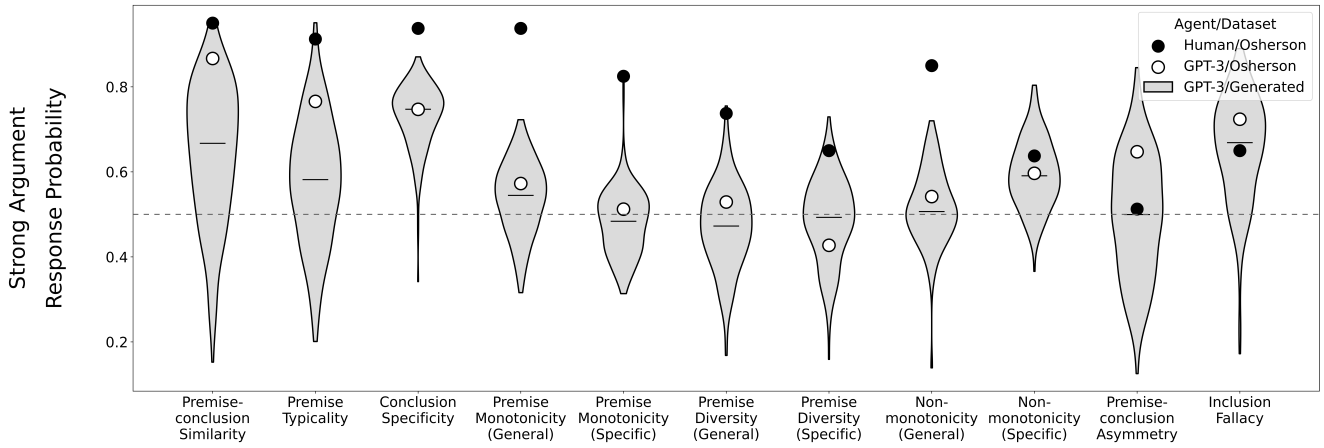

Figure 1: **Inductive reasoning phenomena exhibited by GPT-3 and human reasoners.** Response probability in favour of the stronger of two inductive arguments for the 11 inductive reasoning phenomena shown in Table 1. White dots (GPT-3) and black dots (humans) show response probabilities for the specific argument pairs presented in Osherson et al. (1990), and violin plots (with median shown) reflect GPT-3 responses across all generated argument pairs. While GPT-3 somewhat captures phenomena involving similarity, specificity, and typicality, it performs more poorly on those involving (non)-monotonicity and diversity.

forward to derive which member of the pair is considered stronger by the SCM even without knowing the scores assigned to individual arguments: for Conclusion Specificity, the stronger argument is always the argument to the more specific conclusion, and for Non-Monotonicity the stronger argument is always the one with fewer premises.

For each argument pair we randomly sampled categories as needed, and for the argument pairs based on SCM scores we randomly sampled 2000 argument pairs before picking the 100 with the greatest disparity between their SCM scores. To control for similarity and typicality effects in our set of Premise Diversity and Monotonicity arguments, we considered the strength of the inductive projection (as measured by the SCM) from each individual premise category to the conclusion category. For Premise Diversity, we sampled premise categories such that the second premise category in either argument projected less strongly to the conclusion category than the first premise category. For Monotonicity, we ensured that the third premise category projected less strongly to the conclusion category than at least one of the first two premise categories. Sampling argument pairs in this way ensures that the comparison between strong and weak arguments is driven by diversity or monotonicity respectively, and not by any single premise category in isolation.

The violin plots in Figure 1 summarise the responses of GPT-3 across the arguments sampled for each phenomenon. GPT-3 captures the first four to some extent, and also captures non-monotonicity (specific) and the inclusion fallacy. In all of these cases the median of the violin lies above the dotted 0.5 line, indicating that GPT-3 reliably prefers the stronger argument in each pair. That said, the performance of GPT-3 was more variable and less convincing for phenomena involving Premise Monotonicity and Premise Non-Monotonicity, and it did not capture Premise Diversity at all.

Although GPT-3 does not show a strong effect of premise-conclusion asymmetry, this failure can perhaps be excused because the human data in Figure 1 also reveal no effect (although Osherson et al. (1990) present a second study that does reveal the effect). The results for Premise Diversity, Premise Monotonicity, and Non-monotonicity therefore reveal the greatest limitations of the model. Although all three phenomenena appear to be robust in Western adults, they do not always emerge in other populations (López, Gelman, Gutheil, & Smith, 1992; López et al., 1997). For example, López et al. (1992) found support for similarity, typicality and conclusion specificity in kindergarteners but no evidence for premise diversity and monotonicity, and only partial support for non-monotonicity. Figure 1 therefore raises the possibility that GPT-3 might provide a better account of inductive reasoning in children than adults.

Does GPT-3 account for human argument rankings?

Considering inductive phenomena in isolation is a useful starting point, but this approach is limited because multiple phenomena are relevant to some inferences, and these phenomena sometimes conflict. For example, from the perspective of diversity $\{\text{FLAMINGO, ALBATROSS}\} \rightarrow \text{BIRD}$ is relatively strong because the premise categories are so different from each other. However, it is weak from the perspective of typicality since the premise categories are atypical of birds.

In this section we therefore move beyond the individual phenomena in Table 1 by assessing the ability of GPT-3 to rate the inductive strength of relatively large sets of arguments. Osherson et al. (1990) obtained this data for humans by asking participants to rank two sets of arguments involving mammals. One set included 36 two-premise **Specific** arguments such as $\{\text{COW, CHIMP}\} \rightarrow \text{HORSE}$, where the conclusion in all cases was HORSE. The second included 45 three-premise **General** arguments such as $\{\text{HORSE, COW, MOUSE}\} \rightarrow \text{ALL MAMMALS}$, where the conclusion category

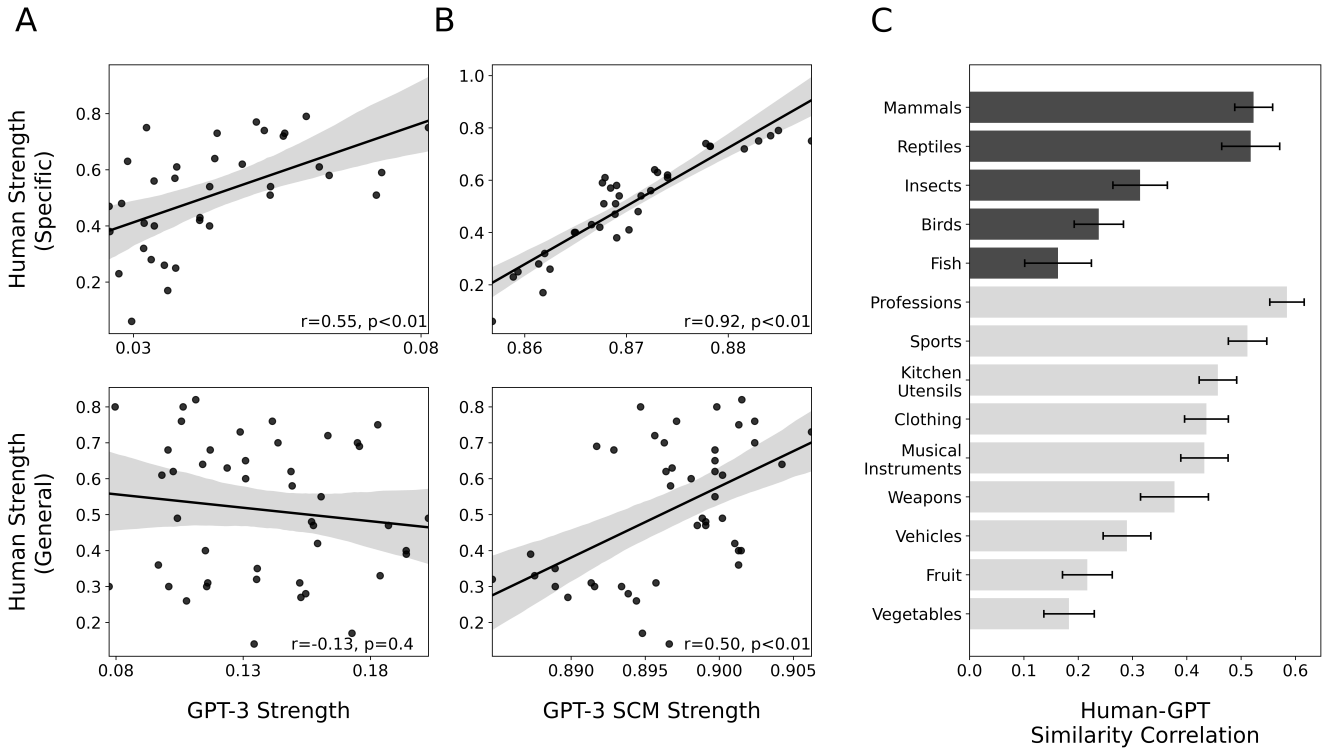

Figure 2: **A**. Overall, the correlation between GPT-3 and human strength ratings for the Specific and General arguments reported in Osherson et al. (1990) is moderate at best. **B**. Correlation between human argument strength ratings and SCM predictions based on GPT-3 derived similarity. Performance is much better, suggesting that the problem with GPT-3 does not lie in the nature of its representations. **C**. Correlations between GPT-3 and human similarity ratings for different categories are moderately strong, again suggesting that the representations of GPT-3 are reasonably accurate. Error bars show standard errors.

was always ALL MAMMALS. For each argument set, we compared mean human rankings with ratings of argument strength elicited from GPT-3 using the method described above.

As Figure 2A reveals, GPT-3 and human argument ratings are moderately correlated for Specific arguments and virtually uncorrelated for General arguments. If anything, the GPT-3 ratings for the general argument set are actually anti-correlated with human responses.

Taken together, our results suggest that GPT-3 performs relatively poorly at capturing human inductive reasoning overall. The model accounts to some degree for six of the 11 qualitative phenomena tested, but the remaining five and the ranking task expose more substantial limitations.

Distinguishing representation from reasoning

Having shown that GPT-3 provides a relatively poor account of human inductive inferences, we now consider two possible explanations for this finding. One possibility is that the internal representations GPT-3 relies on are flawed and do not contain the information necessary to support human-like inductive inferences. A second possibility is that its representations are relatively accurate, but GPT-3 does not use them for inductive inference in the same way that humans do.

We can explore these possibilities by examining the representations that GPT-3 uses. The OpenAI API allows its embeddings to be extracted, allowing us to treat the embedding corresponding to each category label as GPT-3's representation of that category. Each of these representations lies in a 12288-dimension vector space where closeness denotes semantic similarity.[2] The similarity between any two categories according to GPT-3 is therefore calculated as the similarity between the corresponding embeddings. Here we use cosine similarity, but similar results are obtained by using dot product or Euclidean similarity.

We compared these GPT-3 similarity ratings with human similarity ratings reported by De Deyne et al. (2008). Although only the animal categories were relevant to our previous analyses, the full dataset contains 14 superordinate categories; these include clothing, weapons, kitchen utensils, and more. The human ratings we used in our comparison were calculated based on the average similarity rating among the 15-25 participants who rated each category pair.

As Figure 2C shows, the GPT-3 similarity ratings are correlated to some extent with human ratings. This is consistent with previous work suggesting that the internal representations of TLMs can be used to make reasonable predictions about human similarity judgments (Bhatia & Richie, 2021).

[2]These representations are typically derived by combining token embeddings from the hidden layers of the model itself. Although the specific implementation of OpenAI's Embeddings API is not publicly available, it is advertised to be built directly on top of GPT-3's model weights and is thus probably an accurate reflection of its core representation space.

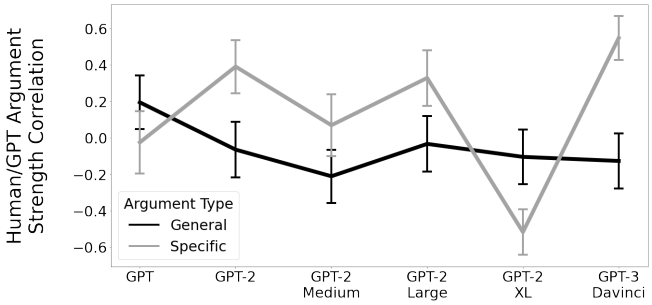

Figure 3: Correlation of different generations of GPT with human argument strength ratings show no consistent improvement in the performance of GPT over time. Error bars show standard errors.

GPT-3 accounts for some superordinate categories better than others, with correlations ranging between 0.16 (fish) and 0.58 (professions). The correlation for mammals is towards the upper end of the range, which suggests that the poor performance of GPT-3 for the mammal-based argument sets in Figure 2A is probably not primarily due to poor representations of mammal categories.

If GPT-3's representations of mammals do capture reliable information, then combining a GPT-3 derived similarity measure with the SCM may provide a relatively good account of human inductive judgments. Figure 2B shows that this hybrid model does indeed account relatively well for the human argument ranking data. The correlation of 0.92 achieved on the specific data set is comparable to the 0.95 correlation achieved when the SCM uses human-generated similarity ratings. The correlation for the general data set is lower (0.49 compared to 0.87 achieved when the SCM uses human similarity ratings), but still substantially higher than the GPT-3 result in Figure 2A.

Will GPT-3 improve with scale?

Our results so far suggest that GPT-3's internal representations may be of sufficient quality to support human-like inferences, but that GPT-3 does not possess a reasoning mechanism that can extract the full value from these representations. Is this limitation fundamental to the design of GPT-3, or is this something that (like many other natural language tasks) we can expect to improve by increasing the size of the model or the quantity of its training data?

To address this question we turn to earlier variants of the GPT family of language models, GPT and GPT-2. They are extremely similar to GPT-3 by design, with their main difference being the scale of their model parameters and training datasets. As GPT variants increase in scale (as measured by model parameter count) by at least one order of magnitude with every generation, leaps in performance across a broad set of language understanding benchmarks have also been observed. If successive generations have improved in their ability to account for human inductive judgments, it seems plausible that this improvement will continue in the future.

To evaluate performance across these generations we used pretrained, off the shelf implementations of previous GPT variants available via the Transformers library (Wolf et al., 2020). We examined five variants in increasing order of scale: GPT, GPT-2 Small, GPT-2 Medium, GPT-2 Large and GPT-2 XL. Each model was given the same prompts and evaluated using the same method described previously.

Figure 3 shows that successive GPT variants failed to demonstrate any clear improvements in how correlated their argument strength ratings were with those of humans. In fact, there seems to be no relationship between scale and performance at all. Uncertainty inevitably remains about the abilities of future variants, but our results provide no reason to think that improvement is simply a matter of scale.

Discussion

We found that GPT-3 provides a relatively poor account of human inductive reasoning, which raises two important directions for future work. First, given that GPT-3 does not closely follow the reasoning principles used by humans and captured by the SCM, how can we understand what GPT-3 is actually doing? A possible way to address this question is to implement a family of interpretable models and to identify which of the models in this family correlate most strongly with GPT-3. We took a preliminary step in this direction by considering a set of variants of the SCM; this includes one that does not include the coverage term and is consistent with the inferences of kindergarteners (López et al., 1992), and another called SumSim (Tenenbaum, Kemp, & Shafto, 2007) that replaces the similarity function used by the SCM with an alternative more consistent with exemplar models of categorization. Because GPT-3 appears to capture similarity and typicality effects but not diversity and non-monotonicity effects, we were optimistic that removing the coverage term from the SCM might yield a model that correlated highly with GPT-3. All of the variants we considered, however, matched GPT-3 relatively poorly, which means that we do not yet have real insight into why GPT-3 reasons as it does.

A second important future direction is to develop computational approaches that maintain the generality and flexibility of GPT-3 – including its ability to handle arguments with non-blank properties – but provide a closer account of human inductive reasoning. Our results exploring the effect of scaling suggest that simply increasing the size of GPT-3 is unlikely to achieve this goal. This means that alternative architectures and/or training objectives will probably be needed. Some researchers discuss intrinsic limitations of large language models: for example, Bender and Koller (2020) suggest that these models are unable in principle to acquire *meanings*, and can succeed only in predicting *forms*. It seems unlikely that the results in this paper expose any such intrinsic limitation, and the respectable performance of the GPT-3/SCM hybrid suggests that a general-purpose model that builds on GPT-3 may be able to perform well on the datasets considered here. Developing such a model is a natural next step towards the ultimate goal of capturing and understanding the rich intricacy of human inductive reasoning.

Acknowledgments

This work was supported in part by the Australian Defence Science & Technology Group (KJR) and ARC Future Fellowship FT19010020 (CK). Code and data are available at https://github.com/S-J-HAN/GPT-3-Property-Induction

References

Bender, E., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *58th Annual ACL*, 5185–5198.

Bhatia, S., & Richie, R. (2021). Transformer networks of human conceptual knowledge. *Psychological Review*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *arXiv:2005.14165*.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030–1048.

Hayes, B. K., & Heit, E. (2018). Inductive reasoning 2.0. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*(3).

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K. J., & Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic Bulletin & Review*, *26*, 1043–1050.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (p. 1725-1744).

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

López, A., Atran, S., Coley, J. D., Medin, D., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, *32*(3), 251–295.

López, A., Gelman, S. A., Gutheil, G., & Smith, E. E. (1992). The development of category-based induction. *Child Development*, *63*(5), 1070–1090.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. (2003). A relevance theory of induction. *Psychonomic Bulletin and Review*, *10*, 517–532.

Misra, K., Ettinger, A., & Taylor Rayz, J. (2021). Do language models learn typicality judgments from text? In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society* (pp. 216–222).

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185–200.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., ... Irving, G. (2021). Scaling language models: Methods, analysis & insights from training Gopher. *arXiv:2112.11446*.

Ransom, K. J., Perfors, A., & Navarro, D. J. (2016). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*, *40*(7), 1775–1796.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*, 665-681.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A Parallel Distributed Processing approach*. Cambridge, MA: MIT Press.

Sap, M., Rashkin, H., Chen, D., LeBras, R., & Choi, Y. (2019). SOCIALIQA: Commonsense reasoning about social interactions. In *Proc. of the 2019 EMNLP-IJCNLP*.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231–280.

Tenenbaum, J. B., Kemp, C., & Shafto, P. (2007). Theory-based Bayesian models of inductive reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: experimental, developmental and computational approaches* (pp. 167–204). Cambridge: Cambridge University Press.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. ACL.