

Peer Review Information

Journal: Nature Human Behaviour

Manuscript Title: Emergent Analogical Reasoning in Large Language Models

Corresponding author name(s): Taylor Webb

Reviewer Comments & Decisions:

Decision Letter, initial version:
--

9th March 2023

Dear Dr Webb,

Thank you once again for your manuscript, entitled "Emergent Analogical Reasoning in Large Language Models," and for your patience during the peer review process.

Your manuscript has now been evaluated by 3 reviewers, whose comments are included at the end of this letter. Although the reviewers find your work to be of high interest, they also raise some important concerns. We are interested in the possibility of publishing your study in Nature Human Behaviour, but would like to consider your response to these concerns in the form of a revised manuscript before we make a decision on publication.

We believe that it will be important to acknowledge and fully discuss the limitation - pointed out by all 3 reviewers - that the LLM, unlike humans, did not have to parse the input visually. However, we do not think it will be necessary to carry out the additional experiments as recommended by Reviewer #1 in their third paragraph. You should also discuss (with appropriate caution) the question of why the current version of GPT performed much better than previous ones, as raised by Reviewer #3,

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

In sum, we invite you to revise your manuscript taking into account all reviewer and editor comments. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

We hope to receive your revised manuscript within 6 weeks. I would be grateful if you could contact us

as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

- Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. When formatting this document, please respond to each reviewer comment individually, including the full text of the reviewer comment verbatim followed by your response to the individual point. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.
- Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Sincerely,
Jamie

Dr Jamie Horder
Senior Editor
Nature Human Behaviour

REVIEWER COMMENTS:

Reviewer #1:
Remarks to the Author:

The manuscript is clearly written and aims to address a very important question of whether zero-shot large language models (LLMs) can readily solve analogical reasoning. Although enjoyed reading the paper, the biggest concern, explained in the following, is about the way that the Digit Matrices (DM) dataset was constructed and evaluated, which is the main effort of this work. The rest of benchmarking looks trivial and notably, there are counterexamples with ChatGPT.

Essentially, the DM problem set is represented via a disentangled and symbolic format that is easy for LLMs to operate with. In Fig1b, the first digit in the cell represents the shape of the object, the second

digit represents the orientation of the rectangular-shape in the middle, and the last digit represents the type of filling pattern. Obtaining such disentangled representations directly from visual sensory inputs is still an open question in machine learning [A], while the DM dataset bypassed this key challenge and provided the symbolic disentangled inputs for LLMs. In other words, the generated DM dataset assumed that a perception system, e.g., a symbolic parser, would provide the symbolic disentangled representations (and has eluded the uncertainty in perception). In fact, it has been argued in [B] that the construction of such appropriate representations is part of the reasoning process that LLMs should take care of (if possible at all). This is not the case in the conducted experiments because the DM format already provided such appropriate representations, making the job of LLMs significantly easier. Please refer to the recent work in [C] for more detailed discussions on this key aspect. This, therefore, casts doubt on why LLMs should be used with such input format where a simple inductive process in 2011 [D] was able to solve Raven's Progressive Matrices with complexities considered in this manuscript e.g., having constant or progression rules.

Consequently, if the full capability of LLMs is of concern, why not use them end-to-end in a modality (i.e., input format) that is also familiar to humans (and not necessarily easier for machines)? Possible directions would be to use vision-based transformers (or other variants) with the original visual Raven's Progressive Matrices, or to "translate" them (without disentangling or making any attempt to solve them) as natural language inputs to LLMs. The latter modality was also suggested in [E] for approaching analogical reasoning problems.

We should also note that we found 3 adults were much slower in solving a DM problem compared to the corresponding Raven (Fig1.b vs Fig1.a). Anyhow, although the DM problems might involve somewhat similar level of problem difficulty as the traditional Raven problems for humans, certainly they will not provide the same level of difficulty as the traditional visual Raven problems for machines.

Minor comment: As observed in the paper, problems with progression are more difficult to solve than no progression problems. Would be great to evaluate more complex rules such as arithmetic plus/mines that have been used in the RAVEN dataset [8].

[A] Greff, van Steenkiste, and Schmidhuber, 'On the Binding Problem in Artificial Neural Networks', arXiv:2012.05208

[B] Chalmers, D. J., French, R. M. & Hofstadter, D. R. High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence* 4, 185–211 (1992).

[C] Hersche, M., Zeqiri, M., Benini, L., Sebastian, A., Rahimi, A. A neuro-vector-symbolic architecture for solving Raven's progressive matrices. arXiv:2203.04571

[D] Rasmussen, D. & EliaSmith, C. A neural model of rule generation in inductive reasoning. *Topics in Cognitive Science* 3, 140–153 (2011).

[E] Shah, V. et al. Knowledge-based analogical reasoning in neuro-symbolic latent spaces. In *Proceedings of the 16th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy)* (2022).

Reviewer #2:

Remarks to the Author:

I have read this paper with great interest. Webb et al. let GPT-3, and in particular, the latest engine text-davinci 3, solve four types of analogy problems: digit matrices, letter-string analogies, four-term verbal analogies, as well as analogies between natural language stories. GPT-3 performed surprisingly well and often times better than humans. However, some differences emerged, for example, in problems that involved some type of physical reasoning.

I overall really enjoyed reading this paper, and I think it would make a great contribution to NHB.

I have --overall-- only three main comments:

1. GPT-3's brittleness:

In their studies, GPT-3 benefited both from the prompt and adapted problem format. Indeed, it is fairly well-known that these models can be very brittle to the particular prompt. I think the claim, therefore, should be that GPT *can* exhibit good reasoning abilities given the right prompt format. Otherwise, I'm sure people can easily show that it doesn't do much at all if the prompt is just changed a little bit.

2. Differences between RPMs and text-based RPMs:

I think it should be discussed a little further that text-based RPMs might be quite different from normal RPMs which require the extraction of features (frequently over-determined) from a visual presentation. I'm mentioning this here because this is also a subject of investigation in AI.

3. Problems appearing in the training set:

I wasn't quite sure about this but how can we be sure that some of the analogical reasoning tasks, for example, Duncker's task have not been in the training set? I actually thought that this was the weakest part of the paper in any case because of this. I liked the matrix problems much more because it's fairly clear GPT has never seen those before.

3. Toning down some of the claims

I think some of the claims of this paper are unnecessarily overblown, for example, "We found that GPT-3 displayed a surprisingly strong capacity for abstract pattern induction, matching or even surpassing human capabilities in most settings." or "nevertheless, the present results indicate that this approach may be sufficient to achieve human-like reasoning abilities, albeit through a radically different route than that taken by biological intelligence".

I think the results are interesting as they are even without fully equating GPT's performance with humans just based on these text-based results.

4. Lack of a discussion on related work:

This was a little disappointing, I thought. The authors make it really sound like they are the only ones working on this. Yet there are many people currently investigating LLM's reasoning abilities.

Most importantly, probably, is the collective effort on the BIG-Bench, which --actually-- contains reasoning tasks already:

<https://arxiv.org/pdf/2206.04615.pdf>

Funnily enough, there's even a task in there by Holyoak et al., as well as Cholet's reasoning corpus. This seems quite similar to what the authors are proposing here so I would at least mention this paper.

Our lab has also worked on GPT-3's reasoning abilities in cognitive tasks that GPT has never seen before. I don't normally ask authors to cite our work (as a reviewer) but here I think it could add to the discussion (and the point about what GPT can and cannot do):

<https://www.pnas.org/doi/10.1073/pnas.2218523120>

Finally, there's a lot of work by Lampinen and colleagues that is also quite related and could be discussed, for example:

<https://arxiv.org/pdf/2207.07051.pdf>

In summary, I very much like this paper and I'm supportive of the approach to using cognitive science tasks to investigate LLMs. I think a modified version of this manuscript can definitely be published in NHB. Yet I think the current article makes unnecessarily big claims and could definitely cite related papers a bit more generously.

I'm happy to sign this review,
Eric Schulz

Reviewer #3:

Remarks to the Author:

This work reports that a recent large language model (LLM) produced by OpenAI solves a wide range of hard analogy problems at a level at or exceeding humans without specific training on these tasks.

Analogy is a core human cognitive ability and was in the past thought to require specific representational forms or bespoke computational architectures. These results challenge that position, since the computational model considered is a generic neural network designed for sequence learning and language processing, rather than analogy per se.

The fact that the ability to resolve analogy problems emerges in such a model is indeed a notable and surprising fact, one of various emergent capacities that have surprised even the developers of such models once they are trained on sufficient data. I share the authors' view that analogy is a critical facet of human intelligence, and the results should have impact in both the cognitive science and AI communities. The experiments carried out are rigorous and principled and this work would constitute a beacon of good scientific practice when compared to a lot of machine learning research.

Major concern

My major concern about the work is that I'm not sure the authors explain what the results tell us about human behaviour or human intelligence. In that sense, the work would seem to fit better in an AI or

Machine Learning forum.

I would like to understand what the results say about theories of how humans learn and/or make analogies, and perhaps about human learning and/or cognition in general. I suspect that the results should say something, and possibly something very profound, but it's hard for me to know exactly what, and I'd like the authors to discuss this (even if exact answers can't be reached).

One difficulty with reaching such conclusions that I can see is that the results show (Fig S2) that a very similar previous model performs substantially worse compared with human performance. Thus, it seems like to understand anything pertaining to humans, it would be necessary to know what the differences are between these two versions of the model. By understanding this difference, we can get insight into what may be critical for human learners to acquire these skills. But the authors don't consider the details of this difference (and indeed, it may not be publicly known).

Less critical concerns

Any parallels between human performance and model performance rests on the results that humans perform similarly on the symbolic versions of the problems that are processed by the LLM and the visual versions that are typically presented to humans. The authors establish this equivalence, but I am a little concerned about the robustness of these findings. I know that the interaction (or lack thereof) between perception and reasoning is a topic of considerable debate in cognitive and neuroscience, and the results seem contingent on there being no such interaction. The authors can't of course be expected to settle this debate, but they should acknowledge that the work rests on this to some extent, and mention the literature/ prior evidence for both positions.

The authors should also mention that undergraduates at top universities is a very biased sample of human populations in general, and the fact that performance on tests like Raven's does in some cases seem to correlate with cultural factors and level of education - so that conclusions about humans in general may be limited. I don't think this fact undermines the interest of the work, but perhaps the scope of the conclusions.

While the authors refer to the fact the model tackles analogies zero-shot, it would help to be more clear about exactly how the problems are fed to the model in the main text of the article (not only the methods section). In a couple of sentences, the authors could state how they feed the tokens into the LLM and the fact that (I assume) they use no prompt whatsoever, merely starting with the first token of the test sample. If they do something other than this, it's even more important to make this clear in the main text.

The authors should state that they will make the experimental stimuli available such that other (particularly AI) researchers can use the evaluation tasks to test future models.

Summary

This is important and timely work and should be published quickly. I am unsure if it is right for Nature Human Behavior and would like to know the thoughts of the authors on my concerns above.

Felix Hill

Author Rebuttal to Initial comments

We would like to thank the reviewers for their many insightful comments and suggestions. We have now revised the manuscript to address these issues. We would also like to note that we have added a number of new results, including:

- 1) A new human behavioral experiment involving letter string analogy problems, together with more extensive testing of GPT-3 in this task domain.
- 2) A new human behavioral experiment involving story analogy problems, combined with a test of GPT-3, as well as an initial investigation of GPT-4.
- 3) Tests of GPT-3 on three additional four-term verbal analogy problem sets, each involving a comparison with human performance.
- 4) More extensive testing of other GPT-3 model variants, including davinci, text-davinci-002, and code-davinci-002.

These additional results make for a much more comprehensive account of GPT-3's emergent analogical capabilities. Below we include a point-by-point reply to the issues raised by the reviewers, along with a description of the corresponding revisions made to the manuscript. The reviewers' comments are presented in blue, and revisions are presented in red. Emphasis (bold text) is occasionally added to the quoted passages, and strikethrough is used to indicate text that has been revised.

Reviewer 1

The rest of benchmarking (beyond the Digit Matrices) looks trivial and notably, there are counterexamples with ChatGPT.

We would like to emphasize that the revised manuscript now includes a substantially expanded set of tests of GPT-3, going well beyond the Digit Matrices dataset that was the primary focus of the original manuscript. These tests include two additional experiments in which we performed a direct comparison between GPT-3 and human participants on both letter string analogy problems and story analogies. The tests also include assessment of GPT-3 on three additional four-term verbal analogy datasets, which, together with the dataset that GPT-3 was tested on in the original manuscript (UCLA VAT), constitute the only four-term verbal analogy datasets for which

human behavioral data is publicly available. The revised manuscript thus covers a very wide range of text-based analogy problems, and provides a comparison with human behavior in each task domain.

It is also important to note that our tests were performed on text-davinci-003, which is distinct from ChatGPT. Our understanding is that these two models have had similar training trajectories, but it is possible that they have a distinct set of abilities when it comes to analogy tasks. The publicly available version of ChatGPT also does not allow control of temperature, which we have found to be an important variable for obtaining optimal performance (we find the best results are obtained with temperature = 0). Accordingly, examples of poor performance in ChatGPT would not strictly speaking be considered counterexamples to the results that we present. We have attempted to clarify the relationship between text-davinci-003 and ChatGPT in the following passage from the revised manuscript (at the beginning of the Results section):

Our evaluation focused on the most recent model variant, text-davinci-003 (referred to simply as ‘GPT-3’ throughout this paper), which was the first to incorporate RLHF (along with the concurrently released, but distinct, ChatGPT model).

Essentially, the DM problem set is represented via a disentangled and symbolic format that is easy for LLMs to operate with. ... Obtaining such disentangled representations directly from visual sensory inputs is still an open question in machine learning [A], while the DM dataset bypassed this key challenge and provided the symbolic disentangled inputs for LLMs. In other words, the generated DM dataset assumed that a perception system, e.g., a symbolic parser, would provide the symbolic disentangled representations (and has eluded the uncertainty in perception). In fact, it has been argued in [B] that the construction of such appropriate representations is part of the reasoning process that LLMs should take care of (if possible at all). This is not the case in the conducted experiments because the DM format already provided such appropriate representations, making the job of LLMs significantly easier. Please refer to the recent work in [C] for more detailed discussions on this key aspect.

[A] Greff, van Steenkiste, and Schmidhuber, ‘On the Binding Problem in Artificial Neural Networks’, arXiv:2012.05208

[B] Chalmers, D. J., French, R. M. & Hofstadter, D. R. High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence* 4, 185–211 (1992).

[C] Hersche, M., Zeqiri, M., Benini, L., Sebastian, A., Rahimi, A. A neuro-vector-symbolic

[architecture for solving Raven's progressive matrices. arXiv:2203.04571](#)

We thank the reviewer for raising this important point. Our revised manuscript addresses this concern in the following ways:

- 1) First, we have taken care to more explicitly acknowledge this limitation of the present results. We agree that GPT-3's success on the Digit Matrices is not equivalent to success on visual RPM problems. We have added the following passages acknowledging this limitation to the final paragraph of the section describing the Digit Matrices results:

One potential concern with the Digit Matrices dataset is that, unlike traditional matrix reasoning problems, the Digit Matrices do not require parsing of visual inputs (object segmentation, disentanglement of visual attributes, etc). GPT-3 is only capable of processing text-based inputs, and therefore cannot be tested directly on visual reasoning...**it will be important in future work to investigate how these reasoning processes might be integrated with visual processing.**

We have also made a number of revisions throughout the paper when referring to the Digit Matrices task. Whereas before we described the Digit Matrices and traditional visual RPM problems as having 'comparable complexity', we are more careful in the revised manuscript to emphasize only that they have the same *rule structure*, or that the Digit Matrices is 'based on' RPM. For example, in the introduction, we now describe the Digit Matrices in the following way:

These tasks included a novel text-based matrix reasoning task ~~with comparable problem structure and complexity as~~ **based on** Raven's Progressive Matrices...

Similarly, in the first paragraph of Results subsection covering the Digit Matrices, we now say the following:

The Digit Matrices dataset ~~is thus comparable~~ **thus has a similar rule structure** to SPM, but is **guaranteed to be novel for both humans and LLMs.**

- 2) Although we certainly agree that the Digit Matrices is not equivalent to visual RPM problems in every respect, the comparable level of performance in human participants for the two tasks does suggest that they involve a common set of core reasoning processes. In the revised manuscript, we are careful to state that it is only these common core processes that we take the Digit Matrices to be measuring. In the final paragraph of the section describing the Digit Matrices results, we now state the

following:

Though there are undoubtedly differences between these two formats, we found that overall error rates for human participants were very similar for the Digit Matrices vs. the original image-based SPM problem set, and showed a similar pattern across problem types (Figure 4). These results suggest that **the two tasks, despite their differences, engage a similar set of core reasoning processes.**

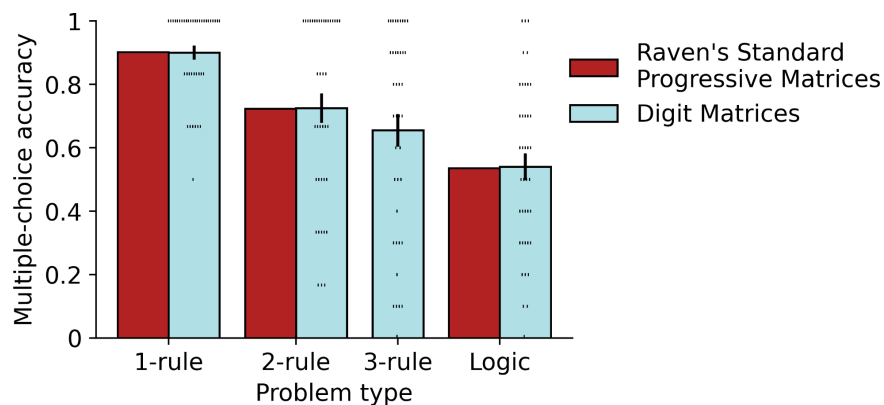


Figure 4: Human performance for Digit Matrices vs. Raven's Standard Progressive Matrices (SPM). SPM [15] does not contain three-rule problems, but performance was very similar across one-rule, two-rule, and logic problems. SPM results reflect average performance for N=80 participants (data from [32]). Digit Matrices results reflect average performance for N=40 participants. Error bars represent standard error of the mean. Each dot represents accuracy for a single participant.

[15] J. C. Raven, *Progressive matrices: A perceptual test of intelligence*, individual form. London: Lewis, 1938.

[32] L. E. Matzen, Z. O. Benz, K. R. Dixon, J. Posey, J. K. Kroger, and A. E. Speed, "Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties," *Behavior research methods*, vol. 42, no. 2, pp. 525–541, 2010.

It is also important to note that, in addition to the results showing similar human performance on the two tasks, it has previously been shown that performance on analogy problems across domains shows a very high degree of correlation. This provides further evidence that there is a core set of reasoning processes that are engaged by analogy problems generally, regardless of whether they are presented in a text-based or visual format. We have added the following note addressing this:

More generally, performance on verbal, visuospatial, and mathematical analogy problems are known to be highly correlated for people [5]

[5] R. E. Snow, P. C. Kyllonen, B. Marshalek, et al., "The topography of ability and learning correlations," *Advances in the psychology of human intelligence*, vol. 2, no. S 47, p. 103, 1984.

- 3) We have added references to the papers noted in the reviewer's comment above [A - C] (Greff et al., Chlamers et al., and Hersche et al.). We thank the reviewer for bringing these papers to our attention.
- 4) We appreciate the reviewer's point that *representation* is an important part of the analogical process, and has been a central concern of cognitive models of analogy. We also agree that our experiments with the Digit Matrices do not clearly address this important aspect of analogical reasoning. Therefore, we performed a new experiment focused on the letter string domain, which was originally developed specifically for the purpose of evaluating how representations and reasoning interact. For instance, in the following example problem:

[a b c d] [a b c e]
[i j j k k l l] [?]

The target (second row), which is presumably initially represented as a series of individual letters, must be *re-represented* as a set of *groups of letters* in order to identify the analogy between these groups and the individual letters in the source (first row). This task allowed us to assess the extent of GPT-3's capacity for re-representation in a text-based format. In the original manuscript, we presented limited results on these problems, but in the revised manuscript we report an entirely new experiment directly comparing human performance and GPT-3. The results (presented in Section 2.2 of the revised manuscript) show that GPT-3 performs comparably to human participants on these tasks, and also shows a similar pattern of performance across different conditions.

These findings suggest that GPT-3 does indeed possess a capacity for re-representation.

Consequently, if the full capability of LLMs is of concern, why not use them end-to-end in a modality (i.e., input format) that is also familiar to humans (and not necessarily easier for machines)? Possible directions would be to use vision-based transformers (or other variants) with the original visual Raven's Progressive Matrices, or to "translate" them (without disentangling or making any attempt to solve them) as natural language inputs to LLMs.

This is an excellent suggestion. We agree that it will be very important in future work to

investigate how the results of the present study can be extended to the visual domain. Unfortunately, GPT-3 itself is only capable of processing text-based input, and so cannot currently be tested in this way. It is likely that future models will be multimodal, which may open up the possibility of directly testing them on visual reasoning tasks. For the time being, given the purely text-based nature of GPT-3, we will take the advice of the Editor and leave this question to be addressed in future work.

Minor comment: As observed in the paper, problems with progression are more difficult to solve than no progression problems. Would be great to evaluate more complex rules such as arithmetic plus/mines that have been used in the RAVEN dataset [8].

This is also an excellent suggestion. We agree that it would be very informative in future work to further investigate the types of rules that GPT-3 is capable of inferring. In the present work, we chose to focus on rules that are commonly represented in the original Raven's Progressive Matrices. Arithmetic rules, as formulated in the RAVEN dataset [1], are, to our knowledge, only present in a single problem from the original Raven's Progressive Matrices – problem 33 of Raven's Advanced Progressive Matrices. This is one of the most difficult problems in the entire set, with an average human accuracy level of ~20% (close to chance performance of 12.5%) [2], which suggests that these rules are extremely difficult for human reasoners to infer. Note also that the rule types commonly referred to as 'figure addition' and 'figure subtraction' are actually identical to the logical rules that are used in the Digit Matrices, and are distinct from the much more difficult arithmetic rules present in RAVEN. Nevertheless, we agree with the reviewer that it would be very interesting in future work to investigate how GPT-3 or other LLMs might perform on more challenging problems, potentially even problems that are too challenging to be solved by most human reasoners.

- [1] Zhang, C., Gao, F., Jia, B., Zhu, Y., & Zhu, S. C. (2019). Raven: A dataset for relational and analogical visual reasoning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5317-5327).
- [2] Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological review*, 97(3), 404.

Reviewer 2

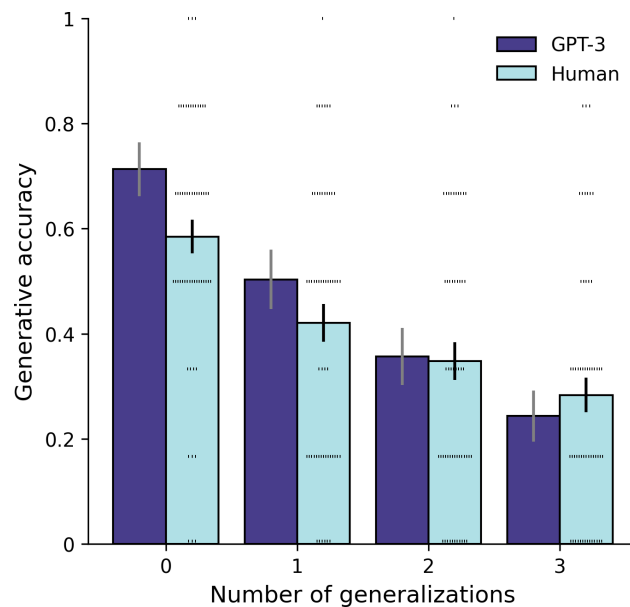
1. GPT-3's brittleness:

In their studies, GPT-3 benefited both from the prompt and adapted problem format. Indeed, it is

fairly well-known that these models can be very brittle to the particular prompt. I think the claim, therefore, should be that GPT *can* exhibit good reasoning abilities given the right prompt format. Otherwise, I'm sure people can easily show that it doesn't do much at all if the prompt is just changed a little bit.

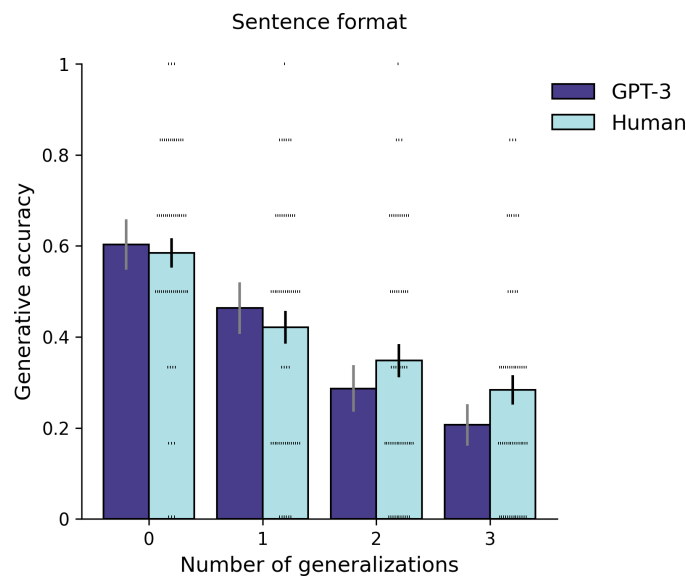
We thank the reviewer for drawing our attention to this important issue. In the original manuscript, we reported that GPT-3's performance on letter string analogy problems was somewhat sensitive to the way in which they were formatted. For instance, GPT-3 did not perform as well when not given a task instruction, or when presented with a problem in the form of a sentence. We have now performed a much more extensive investigation of this issue (presented in section 2.2 of the revised manuscript), including an additional human behavioral experiment, and more extensive testing of GPT-3 on letter string analogies.

These results confirmed our initial findings that GPT-3 is sensitive to formatting for these problems. However, we found that the overall effect of formatting was somewhat less prominent than suggested in the more limited experiments included in the original manuscript. Below is a figure (Figure 6a of the revised manuscript) depicting GPT-3's performance on these problems when formatted in the standard way (using brackets to demarcate the analogs, and with the prompt "Let's try to complete the pattern"):



'Number of generalizations' refers to the number of differences between source and target (grouping in the target, a longer string in the target than the source, etc.). Dots represent results

for each individual human participant. By contrast, below are the results when problems were presented to GPT-3 in the form of a sentence:



Thus, although GPT-3's performance is clearly affected by this change in formatting, it is not a very large effect, and the resulting performance even in the latter case is still on par with human participants in most conditions. Furthermore, the pattern of results across conditions is still very similar for GPT-3 and human participants, with a strong correlation in performance at the level of specific problem types ($r = 0.72$).

Nevertheless, we agree that it is important to fully acknowledge and discuss these aspects of GPT-3's behavior. Accordingly, we have added the following paragraph to the end of the section with letter string results in the revised manuscript:

One important caveat is that GPT-3's performance on this task was somewhat sensitive to the way in which problems were formatted. For instance, performance suffered when no prompt was provided (Figure S5a), or when problems were presented in the form of a complete sentence (Figure S5b). However, even in these cases, GPT-3's zero-shot performance was both within the range of human participants (within one standard deviation), and closely matched the pattern of human performance across problem types (correlation analysis, no prompt: $r(39) = 0.6$, $p < 0.0001$, sentence format: $r(39) = 0.76$, $p < 0.0001$).

Figure S5 is shown below:

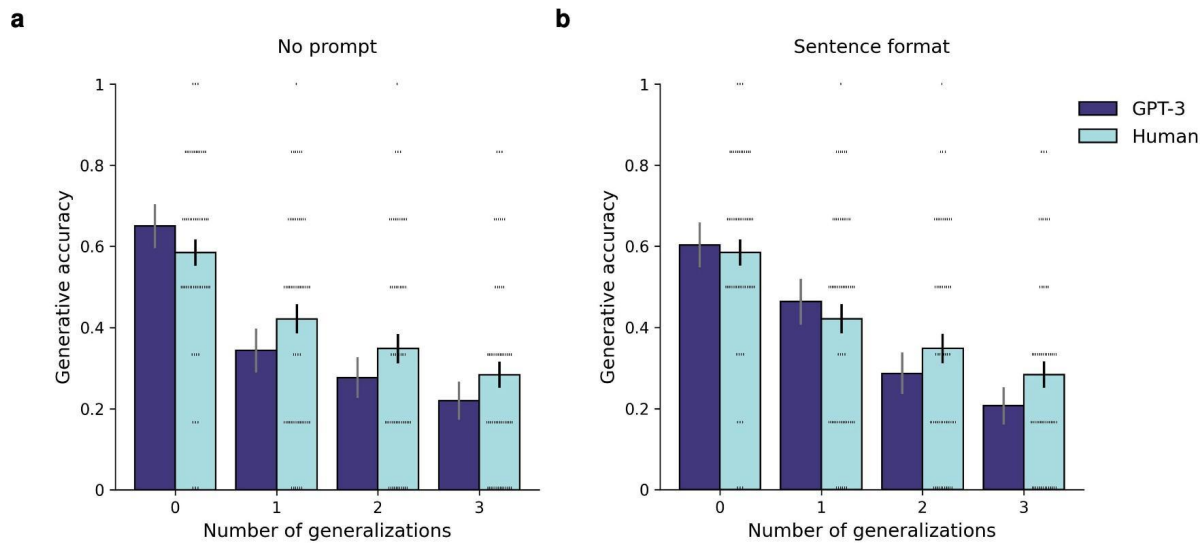


Figure S5: Letter string analogy results with alternative formatting. (a) GPT-3 performed worse when letter string problems were presented without prompt, in comparison with results (shown in Figure 6a) when same problems were formatted in the standard way (using brackets to demarcate the analogs, and with the prompt “Let’s try to complete the pattern”). (b) GPT-3 also performed worse when problems were presented in the form of a sentence (e.g., ‘If a b c d changes to a b c e, then i j k l should change to’). Human results reflect average performance for N=57 participants (UCLA undergraduates). Black error bars represent standard error of the mean across participants. Each dot represents accuracy for a single participant. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems.

2. Differences between RPMs and text-based RPMs:

I think it should be discussed a little further that text-based RPMs might be quite different from normal RPMs which require the extraction of features (frequently over-determined) from a visual presentation.

We agree that it’s important to be clear about the relationship between the Digit Matrices and traditional RPM. We have made an effort in the revised manuscript to clearly discuss this issue. First, we have made sure to remove reference to claims that the Digit Matrices are ‘comparable’ to RPM problems, or that they have ‘similar complexity’, instead emphasizing only that they have a similar rule structure. For example, in the introduction, we now describe the Digit Matrices in the following way:

These tasks included a novel text-based matrix reasoning task ~~with comparable problem~~

~~structure and complexity as based on~~ Raven's Progressive Matrices...

Similarly, in the first paragraph of Results subsection covering the Digit Matrices, we now say the following:

The Digit Matrices dataset ~~is thus comparable~~ **thus has a similar rule structure** to SPM, but is guaranteed to be novel for both humans and LLMs.

We have also added the following paragraph to the end of section where the Digit Matrices results are presented (Section 2.1):

One potential concern with the Digit Matrices dataset is that, unlike traditional matrix reasoning problems, the Digit Matrices do not require parsing of visual inputs (object segmentation, disentanglement of visual attributes, etc). GPT-3 is only capable of processing text-based inputs, and therefore cannot be tested directly on visual reasoning problems. Though there are undoubtedly differences between these two formats, we found that overall error rates for human participants were very similar for the Digit Matrices vs. the original image-based SPM problem set, and showed a similar pattern across problem types (Figure 4). These results suggest that the two tasks, despite their differences, engage a similar set of core reasoning processes. More generally, performance on verbal, visuospatial, and mathematical analogy problems are known to be highly correlated for people [5]. Accordingly, GPT-3's success on the Digit Matrices can be taken as evidence that it has acquired core capabilities underlying analogy, though it will be important in future work to investigate how these reasoning processes might be integrated with visual processing.

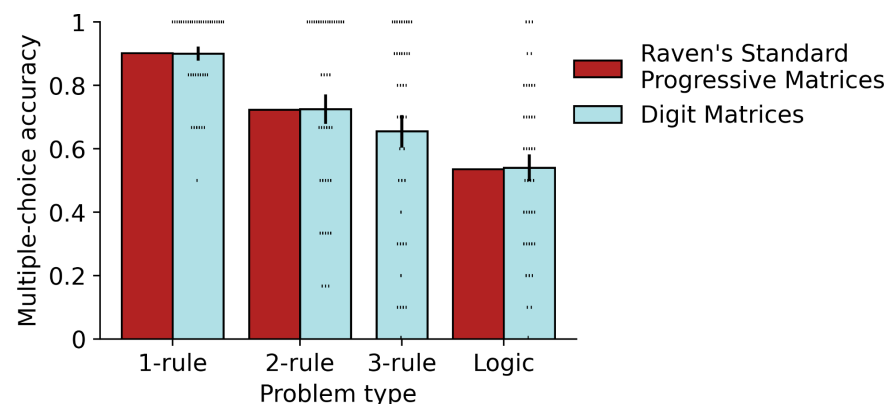


Figure 4: Human performance for Digit Matrices vs. Raven's Standard Progressive Matrices (SPM). SPM [15] does not contain three-rule problems, but performance was very similar across

one-rule, two-rule, and logic problems. SPM results reflect average performance for N=80 participants (data from [32]). Digit Matrices results reflect average performance for N=40 participants. Error bars represent standard error of the mean. Each dot represents accuracy for a single participant.

[5] R. E. Snow, P. C. Kyllonen, B. Marshalek, et al., "The topography of ability and learning correlations," *Advances in the psychology of human intelligence*, vol. 2, no. S 47, p. 103, 1984.

[15] J. C. Raven, *Progressive matrices: A perceptual test of intelligence, individual form*. London: Lewis, 1938.

[32] L. E. Matzen, Z. O. Benz, K. R. Dixon, J. Posey, J. K. Kroger, and A. E. Speed, "Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties," *Behavior research methods*, vol. 42, no. 2, pp. 525–541, 2010.

3. Problems appearing in the training set:

I wasn't quite sure about this but how can we be sure that some of the analogical reasoning tasks, for example, Duncker's task have not been in the training set? I actually thought that this was the weakest part of the paper in any case because of this. I liked the matrix problems much more because it's fairly clear GPT has never seen those before.

We share the reviewer's concern in this issue, and agree that this is a difficult aspect of interpreting results when testing LLMs. We would like to note that, in addition to our novel Digit Matrices task, the letter string analogy experiment in the revised manuscript also features a novel problem set that is guaranteed not to have been in GPT-3's training data. We also believe that GPT-3's performance on the other problem sets that we investigate is not likely to be due to prior exposure, but we agree that it is somewhat less certain in these cases. We agree that it is important to discuss this possibility. We have added the following section to the supplementary material discussing this issue for each of the test materials that we used:

S3. Presence of test materials in GPT-3's training data

Given the massive and uncured nature of GPT-3's training data, it is important to consider the likelihood that our test materials were included in this training data, and the possibility therefore that GPT-3 may have memorized some of these materials (thus undermining their use as a test of zero-shot reasoning).

The Digit Matrices dataset was created specifically for the purposes of our study and therefore certainly was not included in GPT-3's training data. Furthermore, this problem format itself is, to the best of our knowledge, completely novel, and it is thus extremely unlikely that GPT-3 has been

trained on similar problems.

The letter string problem set that we used was also created specifically for the purposes of our study, and so could not have been included in GPT-3's training data. It is possible that GPT-3 has been trained on other letter string analogy problems, as these problems are discussed on a number of webpages (e.g., <https://cogsci.indiana.edu/lap.html>). However, we were not able to obtain any evidence that GPT-3 knows about these problems, e.g., by asking it to describe or give examples of such problems.

The UCLA Verbal Analogy Test [17], and the four-term verbal analogy sets from Sternberg and Nigro [19] and Jones et al. [21] are all available in a downloadable supplement in a recently published paper [40], though to our knowledge they are not published directly in the form of webpages, and are thus unlikely to be included in web crawl data such as the Common Crawl dataset on which GPT-3 was trained. The dataset of SAT problems from Turney et al. [20] is not publicly available, but has been distributed to a relatively small number of researchers. It is possible that any of these datasets were deliberately included in GPT-3's training data (i.e., as a supplement to the web crawl data). However, we were not able to find any evidence that GPT-3 had memorized any of these problems, e.g., by prompting GPT-3 to complete these problems based only on the source analog (the A and B terms alone).

The story analogy materials from Gentner et al. [18] were also included in the downloadable supplement associated with [40], though to our knowledge have not been published directly to any webpages. We were also not able to obtain any evidence that GPT-3 had memorized any of these stories, e.g., by asking it to complete partial stories.

The analogy between the general story and the radiation problem is discussed in the original publication from Gick and Holyoak [22], and is discussed on other webpages as well (e.g., <http://cognitivepsychology.wikidot.com/problem-solving:analogy>). We found evidence that GPT-3 is familiar with this study (it was able to identify the year of publication, and journal in which it was published), but we did not find any evidence that GPT-3 had memorized the details of the analogy itself. GPT-3 was also able to solve the radiation problem via an analogy with lesser known source stories [45]. It seems unlikely that GPT-3's success on these problems is due to prior exposure, but these results should be interpreted with caution. The other materials used to evaluate analogical problem solving are also discussed in the associated publication [23], but we did not find any evidence that GPT-3 is familiar with this study, or that it had memorized these materials.

[17] H. Lu, Y. N. Wu, and K. J. Holyoak, "Emergence of analogy from relation learning,"

Proceedings of the National Academy of Sciences, vol. 116, no. 10, pp. 4176–4181, 2019.

- [18] D. Gentner, M. J. Rattermann, and K. D. Forbus, “The roles of similarity in transfer: Separating retrievability from inferential soundness,” *Cognitive psychology*, vol. 25, no. 4, pp. 524–575, 1993.
- [19] R. J. Sternberg and G. Nigro, “Developmental patterns in the solution of verbal analogies,” *Child Development*, pp. 27–38, 1980.
- [20] P. D. Turney, M. L. Littman, J. Bigham, and V. Shnayder, “Combining independent modules to solve multiple-choice synonym and analogy problems,” in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, 2003, pp. 482–489.
- [21] L. L. Jones, M. J. Kmieciak, J. L. Irwin, and R. G. Morrison, “Differential effects of semantic distance, distractor salience, and relations in verbal analogy,” *Psychonomic bulletin & review*, vol. 29, no. 4, pp. 1480–1491, 2022.
- [22] M. L. Gick and K. J. Holyoak, “Analogical problem solving,” *Cognitive psychology*, vol. 12, no. 3, pp. 306–355, 1980.
- [23] K. J. Holyoak, E. N. Junn, and D. O. Billman, “Development of analogical problem-solving skill,” *Child development*, pp. 2042–2055, 1984.
- [40] N. Ichien, H. Lu, and K. J. Holyoak, “Verbal analogy problem sets: An inventory of testing materials,” *Behavior research methods*, vol. 52, pp. 1803–1816, 2020.
- [45] K. J. Holyoak and K. Koh, “Surface and structural similarity in analogical transfer,” *Memory & cognition*, vol. 15, no. 4, pp. 332–340, 1987.

3. Toning down some of the claims

I think some of the claims of this paper are unnecessarily overblown, for example, “We found that GPT-3 displayed a surprisingly strong capacity for abstract pattern induction, matching or even surpassing human capabilities in most settings.” or “nevertheless, the present results indicate that this approach may be sufficient to achieve human-like reasoning abilities, albeit through a radically different route than that taken by biological intelligence”.

I think the results are interesting as they are even without fully equating GPT's performance with humans just based on these text-based results.

We thank the reviewer for raising this concern. Our revised manuscript addresses this concern in two ways. First, we have added a number of new results in different task domains. These new experiments demonstrate that GPT-3 performs comparably with human participants on letter string analogy problems and four-term verbal analogy problems. We believe that this strengthens our conclusions regarding the generality of GPT-3's analogical capabilities.

However, we agree that it is important to be cautious in interpreting these results. In the revised manuscript, we have made an effort to moderate some of our claims, and also to more thoroughly discuss some of GPT-3's limitations. In particular, the revised manuscript features a new experiment evaluating GPT-3's ability to identify analogies between stories (involving higher-order causal relations), in which we found that GPT-3 showed some limitations relative to human participants. These results are presented in full in Section 2.4 of the revised manuscript, but we present an excerpt below in which we discuss GPT-3's relatively weaker performance on this task:

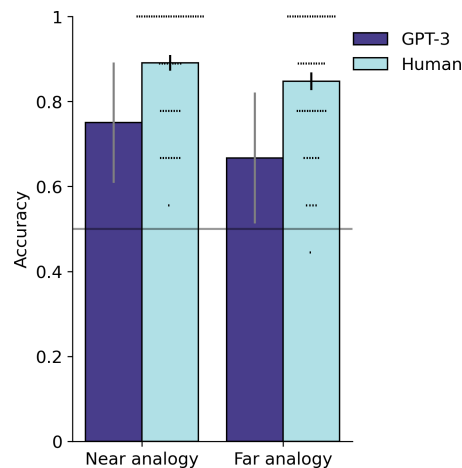


Figure 8: Story analogy results. Results for identification of analogies between stories, using materials from Gentner et al. [18]. When presented with a source story and two target stories, both GPT-3 and human participants showed a preference for target stories that shared higher-order relations with the source vs. those that only shared first-order relations. Near analogy condition involves within-domain comparison between stories with similar entities. Far analogy condition involves cross-domain comparison between stories with different entities. Human results reflect average performance for N=54 participants (UCLA undergraduates). Black error bars represent standard error of the mean across participants. Each dot represents

accuracy for a single participant. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems. Gray horizontal line represents chance performance.

Unlike the other task domains considered in the present work, this was a case in which college students clearly outperformed GPT-3 (logistic regression, main effect of GPT-3 vs. human participants: $OR = 0.37$, $CI = [0.21, 0.63]$, $p = 0.0003$). Indeed, a significant number of participants (15/54) selected the analogous story on every trial.

[18] D. Gentner, M. J. Rattermann, and K. D. Forbus, “The roles of similarity in transfer: Separating retrievability from inferential soundness,” *Cognitive psychology*, vol. 25, no. 4, pp. 524–575, 1993.

Note however that we also evaluated GPT-4 on the story analogy problems, and found that it performed much better:

However, in an initial investigation of GPT-4 [43], we found that it displays stronger performance on this task, more robustly picking the analogous story even in the far analogy condition, and displaying nearly perfect performance in the near analogy condition (Figure S6, Section S4.3).

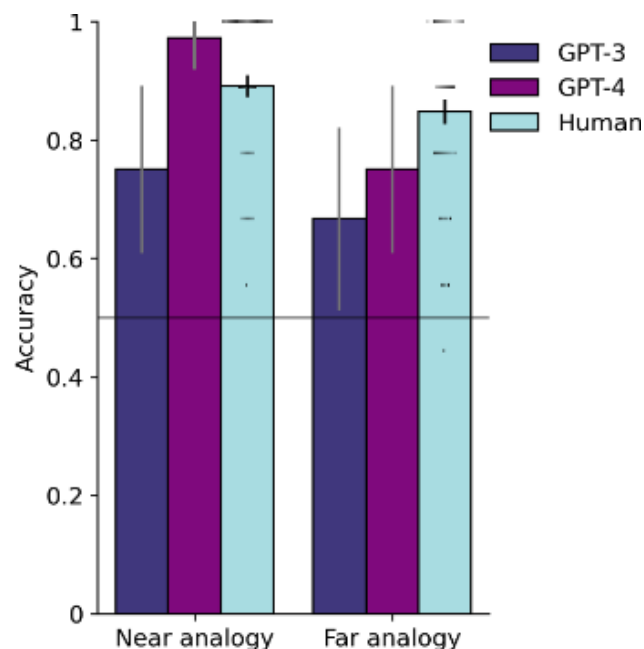


Figure S6: Story analogy results including test of GPT-4. Results for identification of analogies between stories, including both GPT-3 and GPT-4. Whereas GPT-3 was only reliably able to perform the task in the near analogy condition, GPT-4 displayed much more robust performance. GPT-4 showed nearly perfect performance in the near analogy condition, and displayed some sensitivity to higher-order relations in the far analogy condition (binomial test: $p = 0.0039$; null hypothesis is chance-level performance of 0.5). Human results reflect average performance for $N=54$ participants (UCLA undergraduates). Black error bars represent standard error of the mean across participants. Each dot represents accuracy for a single participant. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems. Gray horizontal line represents chance performance.

We have also expanded our consideration of GPT-3's limitations in the discussion section (Section 3):

It is certainly not the case that GPT-3 mimics human analogical reasoning in all respects. **Our tests were limited to processes that can be carried out within a local temporal context, but humans are also capable of retrieving potential source analogs from long-term memory, and ultimately of developing new concepts based on the comparison of multiple analogs.** Unlike humans, GPT-3 does not have long-term memory for specific episodes. It is therefore unable to search for previously-encountered situations that might create useful analogies with a current problem. For example, GPT-3 can use the general story to guide its solution to the radiation problem, but as soon as its context buffer is emptied, it reverts to giving its non-analogical solution to the problem – the system has learned nothing from processing the analogy. GPT-3's reasoning ability is also limited by its lack of physical understanding of the world, as evidenced by its failure (in comparison with human children) to use an analogy to solve a transfer problem involving construction and use of simple tools. GPT-3's difficulty with this task is likely due at least in part to its purely text-based input, lacking the multimodal experience necessary to build a more integrated world model [46]. **Finally, we found GPT-3 was limited in its ability to form analogies based on causal relations, particularly in cross-domain comparisons between stories (far analogy).**

[46] J. L. McClelland, F. Hill, M. Rudolph, J. Baldridge, and H. Schutze, "Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models," *Proceedings of the National Academy of Sciences*, vol. 117, no. 42, pp. 25 966–25 974, 2020.

Finally, we have edited the final sentence of the discussion in the following way:

...the present results indicate that this approach may be sufficient to achieve approximate human-like reasoning abilities...

4. Lack of a discussion on related work:

This was a little disappointing, I thought. The authors make it really sound like they are the only ones working on this. Yet there are many people currently investigating LLM's reasoning abilities.

Most importantly, probably, is the collective effort on the BIG-Bench, which --actually-- contains reasoning tasks already:

<https://arxiv.org/pdf/2206.04615.pdf>

Funnily enough, there's even a task in there by Holyoak et al., as well as Cholet's reasoning corpus. This seems quite similar to what the authors are proposing here so I would at least mention this paper.

Our lab has also worked on GPT-3's reasoning abilities in cognitive tasks that GPT has never seen before. I don't normally ask authors to cite our work (as a reviewer) but here I think it could add to the discussion (and the point about what GPT can and cannot do):

<https://www.pnas.org/doi/10.1073/pnas.2218523120>

Finally, there's a lot of work by Lampinen and colleagues that is also quite related and could be discussed, for example:

<https://arxiv.org/pdf/2207.07051.pdf>

We apologize for this rather significant oversight on our parts. We certainly agree that these other studies are highly relevant to our work, and we agree that it is important to situate our work within the broader literature on emergent reasoning in LLMs. We have now added a reference to the above papers, in addition to a couple other papers that address this general topic. Specifically, we have revised the final sentence of the introduction as follows:

These results add to the growing body of work characterizing the emergent capabilities of LLMs [24–28]...

[24] I. Dasgupta et al., “Language models show human-like content effects on reasoning,” arXiv preprint arXiv:2207.07051, 2022.

[25] A. Srivastava et al., “Beyond the imitation game: Quantifying and extrapolating the

capabilities of language models,” arXiv preprint arXiv:2206.04615, 2022.

[26] J. Wei et al., “Emergent abilities of large language models,” arXiv preprint arXiv:2206.07682, 2022.

[27] S. C. Chan et al., “Data distributional properties drive emergent in-context learning in transformers,” in Advances in Neural Information Processing Systems, 2022.

[28] M. Binz and E. Schulz, “Using cognitive psychology to understand gpt-3,” Proceedings of the National Academy of Sciences, vol. 120, no. 6, e2218523120, 2023.

Reviewer 3

My major concern about the work is that I'm not sure the authors explain what the results tell us about human behaviour or human intelligence. In that sense, the work would seem to fit better in an AI or Machine Learning forum.

I would like to understand what the results say about theories of how humans learn and/or make analogies, and perhaps about human learning and/or cognition in general. I suspect that the results should say something, and possibly something very profound, but it's hard for me to know exactly what, and I'd like the authors to discuss this (even if exact answers can't be reached).

We share the reviewer's sense that these results ought to have (potentially quite significant) implications for our understanding of human behavior and reasoning. We do not currently have any definitive answers to this question. Our goal in the present work was simply to establish, through careful comparison with psychological studies, whether GPT-3 displays analogical capabilities on par with human reasoners, and to what extent it responds similarly to variables that affect problem difficulty in humans. Ultimately, we envision that such models may be useful model systems for understanding the neural basis of human reasoning, just as convolutional networks trained on vision tasks have proven useful for understanding certain aspects of visual brain function. Our ability to investigate this question directly is currently limited by a lack of access to the internal operations of GPT-3. We did, however, identify some (speculative) points of connection between aspects of the transformer architecture and cognitive models of analogy:

A few properties of the transformer architecture [29], on which GPT-3 and other large language models are based, are worth considering here. The first is the central role played by *similarity*. Transformers are built on a self-attention operation, which involves explicitly computing the

similarity between each pair of vectors in the inputs to each layer. This pairwise evaluation of similarity is also a key feature of cognitive models of analogy, where it provides the primary constraint guiding the process of analogical mapping. In traditional symbolic models [50], this takes the form of literal identity between symbols, but in more recent models [51, 52], a graded similarity function that operates over vector-based inputs is used, much like the self-attention operation in transformers. Second, transformer self-attention employs a form of *indirection*, in which one set of embeddings is used to reference another set of embeddings (i.e., keys vs. values) – arguably a form of variable binding. Cognitive scientists have long hypothesized that variable binding plays a central role in analogical reasoning, and abstract reasoning more broadly, as it potentially allows generalization of abstract roles across different contexts [47, 53–57]. It may be that these features of the transformer make it better equipped to perform zero-shot reasoning than other neural architectures. This would also align with recent evidence that the transformer architecture is an important factor contributing toward the emergence of few-shot learning [27].

- [27] S. C. Chan et al., “Data distributional properties drive emergent in-context learning in transformers,” in *Advances in Neural Information Processing Systems*, 2022.
- [50] B. Falkenhainer, K. D. Forbus, and D. Gentner, “The structure-mapping engine: Algorithm and examples,” *Artificial intelligence*, vol. 41, no. 1, pp. 1–63, 1989.
- [51] H. Lu, N. Ichien, and K. J. Holyoak, “Probabilistic analogical mapping with semantic relation networks,” *Psychological Review*, 2022.
- [52] T. W. Webb, S. Fu, T. Bihl, K. J. Holyoak, and H. Lu, “Zero-shot visual reasoning through probabilistic analogical mapping,” *arXiv preprint arXiv:2209.15087*, 2022.
- [53] P. Smolensky, “Tensor product variable binding and the representation of symbolic structures in connectionist systems,” *Artificial intelligence*, vol. 46, no. 1-2, pp. 159–216, 1990.
- [54] K. J. Holyoak and J. E. Hummel, “The proper treatment of symbols in a connectionist architecture,” *Cognitive dynamics: Conceptual change in humans and machines*, vol. 229, p. 263, 2000.
- [55] T. Kriete, D. C. Noelle, J. D. Cohen, and R. C. O’Reilly, “Indirection and symbol-like processing in the pre-frontal cortex and basal ganglia,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 41, pp. 16 390–16 395, 2013.
- [56] T. W. Webb, I. Sinha, and J. D. Cohen, “Emergent symbols through binding in external memory,” in *9th International Conference on Learning Representations, ICLR*, 2021.
- [57] K. Greff, S. Van Steenkiste, and J. Schmidhuber, “On the binding problem in artificial neural networks,” *arXiv preprint arXiv:2012.05208*, 2020.

One difficulty with reaching such conclusions that I can see is that the results show (Fig S2) that a very similar previous model performs substantially worse compared with human performance.

Thus, it seems like to understand anything pertaining to humans, it would be necessary to know what the differences are between these two versions of the model. By understanding this difference, we can get insight into what may be critical for human learners to acquire these skills. But the authors don't consider the details of this difference (and indeed, it may not be publicly known).

We agree that investigating different variants of GPT-3 is a potentially informative route toward understanding the emergence of these phenomena. To better understand this, we carried out more comprehensive tests of different model variants on our analogy tasks, including davinci, code-davinci-002, and text-davinci-002. The results of these tests indicate that emergent analogical reasoning may be a more distributed phenomenon, resulting from multiple factors, rather than a property that emerged specifically in text-davinci-003. We have included a discussion of these results in the supplementary material. This section also mentions our test of GPT-4 on story analogies:

S2. GPT-3 model variants

Since the initial release of GPT-3 [13], OpenAI has released a number of updates to the original base model. The largest version (175B parameters) of the base model, davinci, was trained exclusively on next-token prediction using a web-based corpus of text data. Code-davinci-002 was further trained on next-token prediction using a dataset of publicly available code from GitHub [30]. Text-davinci-002 and text-davinci-003 were both fine-tuned to respond appropriately to prompts [31]. Text-davinci-002 was initialized with code-davinci-002, and then fine-tuned using supervised learning based on a set of example prompts and responses. Text-davinci-003 was initialized with text-davinci-002, and then further fine-tuned using reinforcement learning from human feedback (RLHF). In RLHF, a reward model (a separate neural network) is first trained to predict human ratings for pairs of human-generated prompts and language-model responses, and this reward model is then used to fine-tune the primary language model through reinforcement learning. More details on the different model variants and training objectives can be found at <https://platform.openai.com/docs/model-index-for-researchers>.

We evaluated all four of these variants on Digit Matrices (Figure S1), letter string analogies (Figure S2), and four-term verbal analogies (Figure S3). Text-davinci-003 displayed the best overall performance, but other model variants performed well on a subset of tasks. For instance, code-davinci-002 performed well on the Digit Matrices and letter string problems. These task domains both involve simple alphanumeric characters and highly regular relational structure, similar to computer code. It therefore seems likely that code-davinci-002's strong performance on

these tasks was a consequence of having been trained on code. By contrast, code-davinci-002 performed very poorly on four-term verbal analogies (near chance performance), whereas the original davinci model performed relatively well on these problems. This suggests that code-davinci-002's ability to model synthetic code-like structures may have come at the cost of the ability to process more real-world relational concepts. Text-davinci-002, and especially text-davinci-003, seem to have combined both of these abilities, perhaps as a result of prompt training, though these models may have also received additional fine-tuning on the original language modeling task [31]. Finally, it seems likely that prompt training improved text-davinci-002 and text-davinci-003's ability to perform tasks without the need for few-shot task demonstrations, therefore making it easier to evaluate these models' latent capabilities in a zero-shot setting.

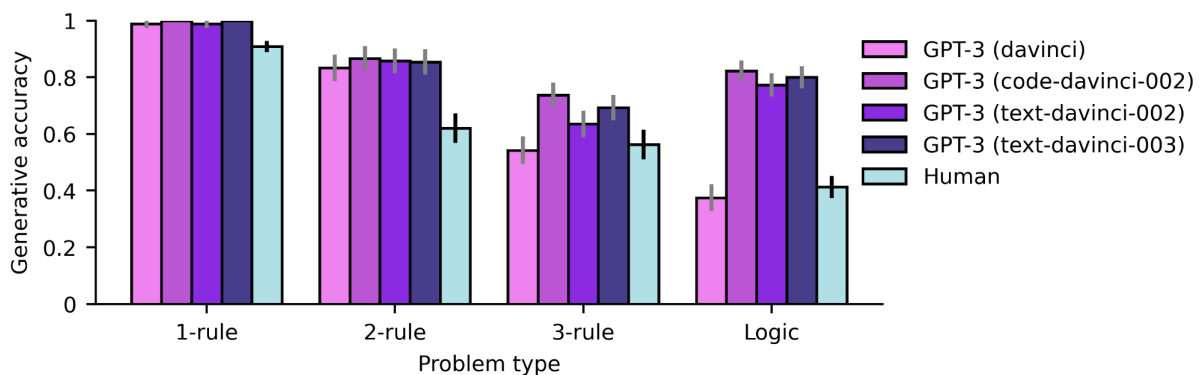


Figure S1: Matrix reasoning results for all GPT-3 variants. Zero-shot results on Digit Matrices for four GPT-3 model variants: davinci, code-davinci-002, text-davinci-002, and text-davinci-003. Results reflect generative accuracy for major problem types, including transformation problems with between one and three rules, and logic problems. Human results reflect average performance for N=40 participants. Black error bars represent standard error of the mean across participants. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems.

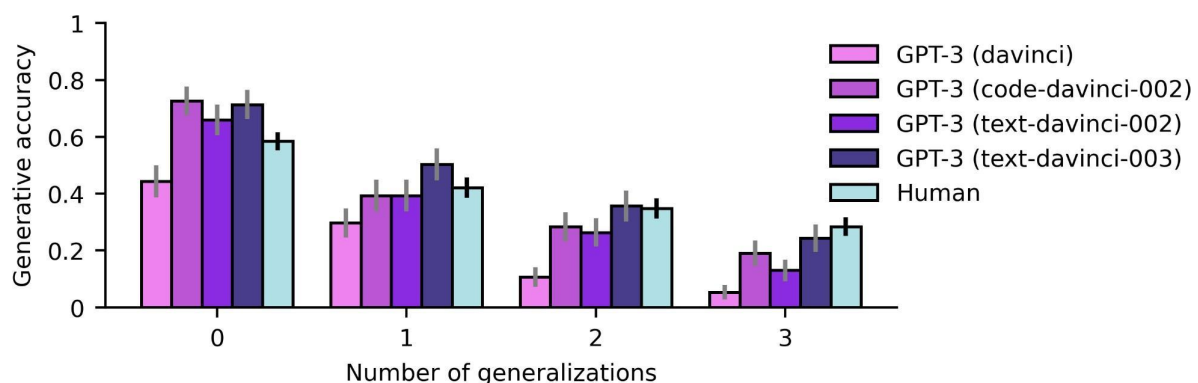


Figure S2: Letter string analogy results for all GPT-3 variants. Letter string analogy results for four GPT-3 model variants: davinci, code-davinci-002, text-davinci-002, and text-davinci-003. Results reflect generative accuracy as a function of the number of generalizations between source and target. Human results reflect average performance for N=57 participants. Black error bars represent standard error of the mean across participants. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems.

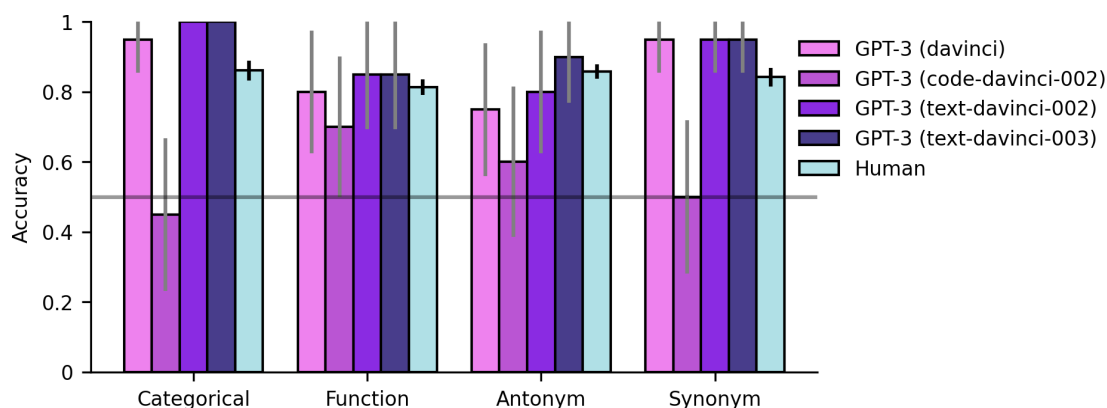


Figure S3: Four-term verbal analogy results for all GPT-3 variants. Results on UCLA Verbal Analogy Test (VAT) for four GPT-3 model variants: davinci, code-davinci-002, text-davinci-002, and text-davinci-003. Results reflect multiple-choice accuracy for problems involving different relation categories. Gray horizontal line represents chance performance. Human results reflect average performance for N=57 participants. Black error bars represent standard error of the mean across participants. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems.

We also performed an initial investigation of GPT-4 on the story analogy problems from Gentner et al. [18] (Figure S6). GPT-4 showed significant improvement on this task relative to GPT-3, more reliably identifying the target story that shared higher-order relations with the source, and providing more precise explanations (Section S4.3). We were not able to test GPT-4 on the other analogy tasks due to a lack of API access. Very little is known about the details of GPT-4, but it is likely that this improvement stems at least in part from increased scale of both model and training set [43].

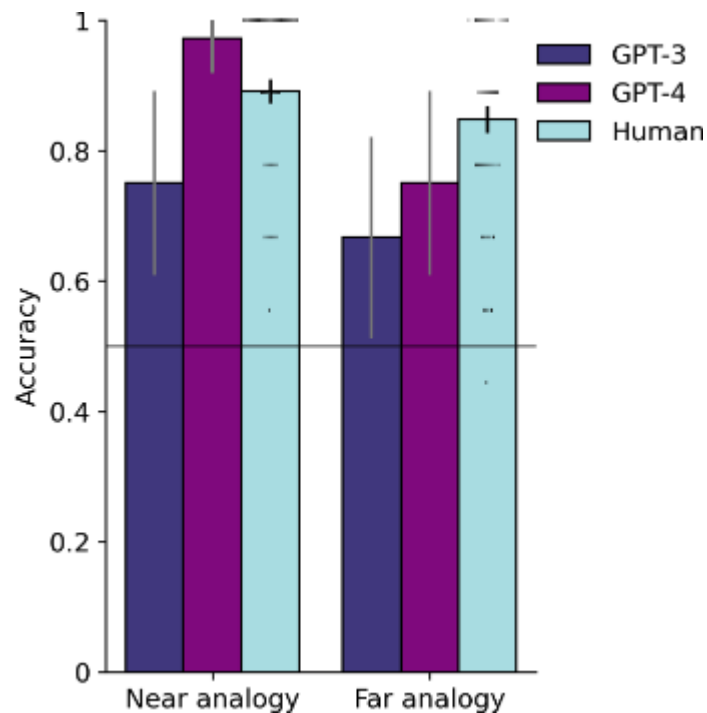


Figure S6: Story analogy results including test of GPT-4. Results for identification of analogies between stories, including both GPT-3 and GPT-4. Whereas GPT-3 was only reliably able to perform the task in the near analogy condition, GPT-4 displayed much more robust performance. GPT-4 showed nearly perfect performance in the near analogy condition, and displayed some sensitivity to higher-order relations in the far analogy condition (binomial test: $p = 0.0039$; null hypothesis is chance-level performance of 0.5). Human results reflect average performance for $N=54$ participants (UCLA undergraduates). Black error bars represent standard error of the mean across participants. Each dot represents accuracy for a single participant. Gray error bars represent 95% binomial confidence intervals for average performance across multiple problems. Gray horizontal line represents chance performance.

We have also revised the description of GPT-3 in the main text as follows (beginning of results):

GPT-3 is a large-scale (175B parameters), transformer-based [29] language model developed by OpenAI [13]. The original base model was trained on a web-based corpus of natural language consisting of over 400 billion tokens, using a training objective based on next-token prediction (given a string of text, the model is trained to predict the token that will most likely appear next). A number of variants on this base model have since been developed by fine-tuning it in various ways. These include training the model to generate code [30], and

training it to respond appropriately to human prompts, using either supervised learning or, most recently, reinforcement learning from human feedback (RLHF) [31]. Our evaluation focused on the most recent model variant, text-davinci-003 (referred to simply as ‘GPT-3’ throughout this paper), which was the first to incorporate RLHF (along with the concurrently released, but distinct, ChatGPT model). We found that text-davinci-003 displayed particularly strong performance on our analogy tasks, but earlier model variants also performed well in some task settings, suggesting that multiple factors contributed to text-davinci-003’s analogical capabilities (Figures S1- S3). See Section S2 for further discussion.

[13] T. Brown et al., “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[30] M. Chen et al., “Evaluating large language models trained on code,” *arXiv preprint arXiv:2107.03374*, 2021.

[31] L. Ouyang et al., “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, 2022.

Any parallels between human performance and model performance rests on the results that humans perform similarly on the symbolic versions of the problems that are processed by the LLM and the visual versions that are typically presented to humans. The authors establish this equivalence, but I am a little concerned about the robustness of these findings. I know that the interaction (or lack thereof) between perception and reasoning is a topic of considerable debate in cognitive and neuroscience, and the results seem contingent on there being no such interaction. The authors can’t of course be expected to settle this debate, but they should acknowledge that the work rests on this to some extent, and mention the literature/ prior evidence for both positions.

We agree that the original manuscript did not devote sufficient attention to this issue. Our view is that interactions between perceptual and reasoning processes (often dubbed *re-representation* in the analogy literature) actually do play an important role in analogy. Since our Digit Matrices task did not provide a clear opportunity to investigate this issue, we performed a new experiment comparing GPT-3 and human performance on letter string analogies, which were originally designed in part to investigate the issue of re-representation. GPT-3 performed comparably to human participants on these problems, providing evidence that it does possess re-representational capacities (results presented in Section 2.2 of the revised manuscript).

The authors should also mention that undergraduates at top universities is a very biased sample of human populations in general, and the fact that performance on tests like Raven’s does in some cases seem to correlate with cultural factors and level of education - so that conclusions

about humans in general may be limited. I don't think this fact undermines the interest of the work, but perhaps the scope of the conclusions.

We share this concern, and accordingly have taken care to emphasize throughout the manuscript that our results apply specifically to college students. For instance, in the introduction, we have revised the following description of our Digit Matrices results:

...we found that GPT-3 performed as well or better than ~~humans~~ **college students** in most conditions...

And in the section describing another new experiment involving story analogies (Section 2.4 of the revised manuscript:

Unlike the other task domains considered in the present work, this was a case in which **college students** clearly outperformed GPT-3...

While the authors refer to the fact the model tackles analogies zero-shot, it would help to be more clear about exactly how the problems are fed to the model in the main text of the article (not only the methods section). In a couple of sentences, the authors could state how they feed the tokens into the LLM and the fact that (I assume) they use no prompt whatsoever, merely starting with the first token of the test sample. If they do something other than this, it's even more important to make this clear in the main text.

We agree that it is important to be clear about this. We have now added the following statement to the section with results for the Digit Matrices (Section 2.1):

Digit Matrices problems were presented to GPT-3 without any prompt or in-context task examples.

We have also added the following statement to the section with results for letter string analogies (Section 2.2), for which we did use a prompt:

Problems were presented to GPT-3 along with a prompt ('Let's try to complete the pattern:'), using a format similar to the Digit Matrices.

And the following description is now given in the section describing the results for four-term analogy problems (Section 2.3):

For each problem, GPT-3 was evaluated by presenting the problem together with each potential answer choice, and selecting the option for which GPT-3 assigned a higher log probability. The problem and GPT-3's choice were then appended to the context window for the next problem. This was done to simulate any contextual effects that might arise when solving multiple problems in a row, as human participants typically do.

The authors should state that they will make the experimental stimuli available such that other (particularly AI) researchers can use the evaluation tasks to test future models.

We have now added the following statement on data availability:

Data Availability

Data for all human behavioral experiments, along with the Digit Matrices, letter string analogy, and UCLA VAT problem sets, can be downloaded from:

[https://github.com/taylorwebb/emergent analogies LLM](https://github.com/taylorwebb/emergent_analogies_LLM)

Decision Letter, first revision:

23rd May 2023

Dear Dr. Webb,

Thank you for submitting your revised manuscript "Emergent Analogical Reasoning in Large Language Models" (NATHUMBEHAV-22123335A). It has now been seen by the original referees and their comments are below. I apologize for the delay in the review process.

Please note that Reviewer #2 did not submit a review. Reviewer #3 did not leave any comments visible to the authors, but in comments to the editors, they said they are happy with the manuscript and recommend publication.

As you can see, the reviewers find that the paper has improved in revision. We will therefore be happy in principle to publish it in Nature Human Behaviour, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions!

Sincerely,
Jamie

Dr Jamie Horder
Senior Editor
Nature Human Behaviour

Reviewer #1 (Remarks to the Author):

I really appreciated all improvements and edits made in the revised manuscript. However, I believe the authors are still shy to explicitly clarify and describe the fundamental differences between Digit Matrices and Raven's Progressive Matrices in various parts of the manuscript. This could create confusion for readers that will just scan through the abstract and intro. As suggested by reviewer #2, it is important to tone down some of the claims made in manuscript. For text-based Digit Matrices, it can be done by a crystal-clear explanation and putting it in contrast with the visual-based Raven's Progressive Matrices. Here are some parts in the manuscript, with some suggestions, that can be clarified accordingly:

- Abstract: "including a novel text-based matrix reasoning task closely modeled on Raven's Progressive Matrices." --> a novel text-based matrix reasoning task that emulates a particular set of Raven's Progressive Matrices without involving visual perception.
- Intro: "These tasks included a novel text-based matrix reasoning task based on Raven's Progressive Matrices" --> These tasks included a novel text-based matrix reasoning task by eluding the uncertainty in visual perception from a particular set of Raven's Progressive Matrices.
- In Fig1. "a novel text-based matrix reasoning task designed to emulate Raven's Standard Progressive Matrices" --> a novel text-based matrix reasoning task designed to emulate a particular set of rule structures in Raven's Progressive Matrices by providing symbolic disentangled representations of the visual objects in the panels.
- In Section 2.1. The following modification is not accurate, hence it requires more elaboration: "These results suggest that the two tasks, despite their differences, engage a similar set of core reasoning processes." --> These results suggest that the two tasks, despite their differences in accessing symbolic disentangled inputs, engage a similar set of core reasoning processes as long as the four types of rules are of concern.

Author Rebuttal, first revision:

We would like to thank reviewer 1 for their continued engagement and additional suggestions for improving the paper. We have now revised the paper to address these remaining concerns. Specifically, the reviewer suggested that we should more carefully characterize the relationship between our Digit Matrices dataset and traditional visual matrix reasoning

problems. We agree with this sentiment, and have revised the paper in the following two ways:

- 1) We have more explicitly stated that the Digit Matrices, unlike traditional visual matrix reasoning problems, do not require disentangled representations to be derived from pixel-level inputs, and instead only test for the ability to induce abstract rules from representations that are already disentangled.
- 2) We note that the Digit Matrices are based specifically on the rule structure of the Raven's Standard Progressive Matrices (SPM) problem set, and do not incorporate all rule types found in other similar problem sets, such as Raven's Advanced Progressive Matrices, or datasets in the machine learning literature (e.g., PGM or RAVEN). We have made sure to revise our claims to clarify that our comparison is specifically with SPM.

The following excerpts illustrate the revisions that have been made to address these issues (bold is used to indicate new text):

Abstract:

Here, we performed a direct comparison between human reasoners and a large language model (the text-davinci-003 variant of GPT-3) on a range of analogical tasks, including a **non-visual** matrix reasoning task **based on the rule structure of Raven's Standard Progressive Matrices**.

Intro:

These tasks included a novel text-based matrix reasoning task based on **the rule structure of Raven's Standard Progressive Matrices (SPM)**, a visual analogy problem set commonly viewed as one of the best measures of fluid intelligence. **Unlike the original visual SPM problems, our Digit Matrices task was purely text-based so that it could be used to evaluate GPT-3's ability to induce abstract rules (though not the ability to do so directly from visual inputs).**

Figure 1:

Matrix reasoning results show average accuracy on all problems in Digit Matrices problem set, a novel text-based matrix reasoning task designed to emulate Raven's **Standard Progressive Matrices (SPM)** problems. **Note that the Digit Matrices were purely text-based, and therefore do not test for the ability to perform abstract reasoning directly over visual**

inputs, as in the original SPM.

Section 2.1:

It is important to highlight the differences between the Digit Matrices and traditional visual matrix reasoning problems. In order to solve visual matrix reasoning problems, pixel-level inputs must be parsed into objects, and visual attributes (shape, size, etc.) must be disentangled. In the Digit Matrices, the text-based inputs are already parsed and disentangled, essentially providing GPT-3 (which is not capable of visual processing) with pseudo-symbolic inputs. Interestingly, despite these significant differences, we found that overall error rates for human participants were very similar for the Digit Matrices vs. the original image-based SPM problem set, and showed a similar pattern across problem types (Figure 4). These results suggest that, while the Digit Matrices do not engage the visual processes involved in traditional SPM problems (i.e., deriving disentangled representations from pixel-level inputs), they likely engage a similar set of core reasoning processes (i.e., inducing abstract rules from those representations).

Final Decision Letter:

Dear Dr Webb,

We are pleased to inform you that your Article "Emergent Analogical Reasoning in Large Language Models", has now been accepted for publication in *Nature Human Behaviour*.

Please note that *Nature Human Behaviour* is a Transformative Journal (TJ). Authors whose manuscript was submitted on or after January 1st, 2021, may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. IMPORTANT NOTE: Articles submitted before January 1st, 2021, are not eligible for Open Access publication. [Find out more about Transformative Journals](#)

Authors may need to take specific actions to achieve [compliance](#) with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to [Plan S principles](#)) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including [self-archiving policies](#). Those licensing terms

will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see <http://www.nature.com/nathumbehav/info/gta>). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

An online order form for reprints of your paper is available at <https://www.nature.com/reprints/author-reprints.html>. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Human Behaviour as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Human Behaviour logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

We look forward to publishing your paper.

With best regards,

Jamie

Dr Jamie Horder
Senior Editor
Nature Human Behaviour