

Response: Emergent analogical reasoning in large language models

Damian Hodel¹ and Jevin West¹

¹Center for an Informed Public, Information School, University of Washington
hodeld@uw.edu, jevinw@uw.edu

August 2023

1 Introduction

In their recent Nature Human Behaviour paper, "Emergent analogical reasoning in large language models," (Webb, Holyoak, and Lu, 2023) the authors argue that "large language models such as GPT-3 have acquired an emergent ability to find zero-shot solutions to a broad range of analogy problems." In this response, we provide counterexamples of the letter string analogies. In our tests, GPT-3 fails to solve even the easiest variants of the problems presented in the original paper.

Given the public's attention on large language models (LLMs) and the hype surrounding their capabilities and this paper in particular^{1,2,3}, we felt it is important to respond and show where the results do not hold. LLMs may have the ability to solve some analogy problems, but stronger evidence is needed, in particular for the claimed zero-shot analogical reasoning. During this hype phase of AI, it is important that we examine other mechanisms for one-off successes (e.g., memorization from trained data).

Others have commented on this paper and want to note these contributions. Mitchell (2023) discusses this paper, focusing on the letter string and digit matrix analogy problems. Mitchell disagrees that "the digit matrix problems are essentially equivalent in complexity and difficulty to Ravens Progressive Matrix problems." Further, Mitchell presents individual counterexamples of the letter string problems where GPT-3 makes nonhuman-like errors as evidence against

¹<https://www.tagesanzeiger.ch/beherrscht-die-kuenstliche-intelligenz-analogien-768020720454>

²<https://www.news-medical.net/news/20230731/AI-language-model-GPT-3-performs-about-as-well-as-college-undergraduates-in-analogical-reasoning.aspx>

³<https://www.sciencemediacentre.org/expert-reaction-to-study-looking-at-gpt-3-large-language-model-and-ability-to-reason-by-analogy/>

the claimed robustness of GPT-3 in analogy reasoning. Our results concur with Mitchell’s conclusion. Mitchell also points out that the term ”accuracy” implies that there was only one correct answer to each problem, which is an assumption implicitly made by the authors of the original paper. Following Mitchell (2023), we adopt Webb, Holyoak, and Lu (2023)’s assumption in our paper and use the same terms, i.e. ”accuracy” and ”performance.”

2 Methods

Original transformation types					
Extend sequence		Successor		Predecessor	
a b c d	→ a b c d e	a b c d	→ a b c e	b c d e	→ a c d e
i j k l	→ i j k l m	i j k l	→ i j k m	i j k l	→ h j k l
Remove redundant letter		Fix alphabetic sequence		Sort	
a b b c d e	→ a b c d e	a b c w e	→ a b c d e	a d c b e	→ a b c d e
i j k k l m	→ i j k l m	i j k x m	→ i j k l m	k j m l i	→ i j k l m
Modified transformation types					
Extend sequence		Successor		Predecessor	
a b c d	→ a b c d f	a b c d	→ a b c f	c d e f	→ a d e f
i j k l	→ i j k l n	i j k l	→ i j k n	j k l m	→ h k l m
Remove redundant letter		Fix alphabetic sequence		Sort	
a c e g i i	→ a c e g i	a c e g o	→ a c e g i	k f a p u	→ a f k p u
i k k m o q	→ i k m o q	i k x o q	→ i k m o q	i m k o q	→ i k m o q
Modified transformation types with synthetic alphabet					
Synthetic alphabet					
x y l k w b f z t n j r q a h v g m u o p d i c e					
Extend sequence		Successor		Predecessor	
x y l k	→ x y l k b	x y l k	→ x y l b	l k w b	→ x k w b
t n j r	→ t n j r a	t n j r	→ t n j a	n j r q	→ z j r q
Remove redundant letter		Fix alphabetic sequence		Sort	
x l w w f t	→ x l w f t	x l w r t	→ x l w f t	x l f w t	→ x l w f t
t t j q h g	→ t j q h g	t j p h g	→ t j q h g	j t q h g	→ t j q h g

Figure 1: Letter string analogies and their transformations for the original paper and this response. We introduce a synthetic alphabet into the task and apply two types of letter sequence modifications, both based on increasing the interval from one to two letters. For the transformation types ’extend sequence’, ’successor’, and ’predecessor’, the modification only affects the *letter to change* (last or first letter). For ’remove redundant letter’, ’fix alphabetic sequence’, and ’sort’, the interval is increased for the complete letter sequence. We apply the same modifications to the problems generated with the synthetic alphabet.

In order to test GPT-3’s generality in zero-shot analogical reasoning, we modify the letter string analogies and compare GPT-3’s performance between the real alphabet and a synthetic alphabet, respectively. Figure 1 shows examples of the original and the modified letter string analogy problems. We create the synthetic alphabet by randomly changing the order of the letters of the real alphabet. If the claim regarding zero-shot is true, we can expect similar performance on both alphabets. To feed the synthetic alphabet to GPT-3, we modify the pattern of the original prompt as follows:

Use this fictional alphabet: [a b c d e f g h i j k l m n
o p q r s t u v w x y z]. Let’s try to complete the pattern:

[a b c d] [a b c d j]

[i j k l] [

We apply this change in both alphabet settings to control for the effects of the prompt format.

To rule out the possibility that GPT-3 merely replicates the fed sequence of letters, we increase the size of the interval from one to two letters. We do this in two ways. For the problem types ‘extend sequence’, ‘successor’, and ‘predecessor’, we increase the interval size for the *letter to change* from one to two. For the problem types ‘remove redundant letter’, ‘fix alphabetic sequence’, and ‘sort’, we increase the interval size of the *complete letter sequence* from one to two⁴. Analogous to the prompt modification, we apply the change in the interval to both alphabet settings.

We report the results for four settings: the original tasks as reported in Webb, Holyoak, and Lu, 2023, modified letter string tasks, modified tasks including the modified prompt, and modified tasks on the synthetic alphabet. Our code for reproducing Figure 2 is available on Github⁵. For each problem type, we create 50 instances to mirror the original paper. Using code from the original paper, we replicated the evaluation and analysis conducted in the original paper. The settings are as follows: model variant=text-davinci-003, temperature=0, maximum length=20.

3 Results

Figure 2 shows the average performance of the original and modified letter string problems with N=50 instances for each transformation type. The modifications

⁴It is worth noting that we apply this modification to both the source (the first row for each example in Figure 1) and the target (the second row for each example in Figure 1), minimizing the difficulty of the modified problems and allowing us to compare our tests to the zero-generalization problems given in the original paper.

⁵https://github.com/hode1d/emergent_analogies_LLM_fork

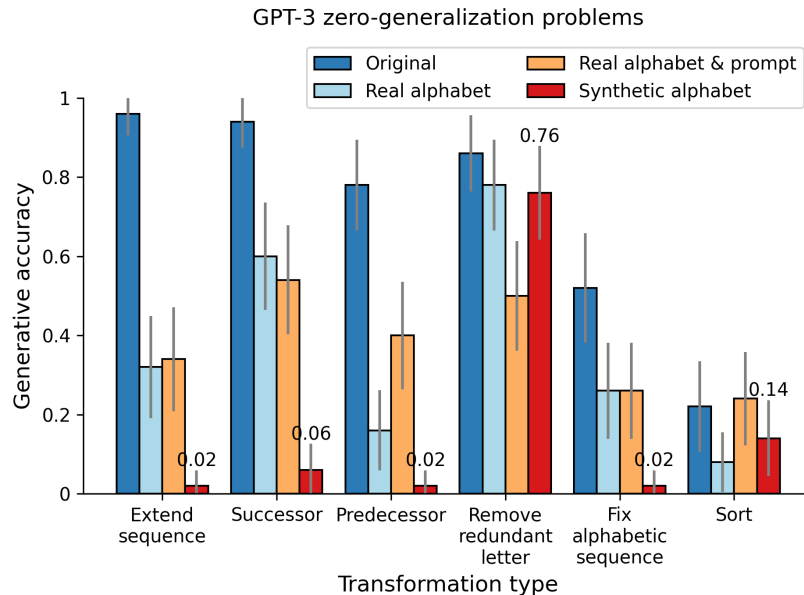


Figure 2: GPT-3 performance for zero generalization letter string problems for the original experiment (dark blue) and with the larger interval size (light blue), larger interval size with alphabet prompt (orange), and larger interval size with synthetic alphabet (red). The figure and the order of the transformation types correspond to Figure 6a in the original paper. These results reflect an average performance for $N=50$ instances.

include a larger interval size and the synthetic alphabet⁶. There are three main takeaways from the results. (1) The introduction of a synthetic alphabet, as a way of testing zero-shot reasoning, drastically decreases the generative accuracy. For the modified problems 'extend sequence', 'successor' or 'predecessor', and 'fix alphabetic sequence', the generative accuracy is close to zero (< 0.1) for the synthetic alphabet. Only for 'remove redundant letter' and 'sort' does GPT-3 achieve accuracy in a similar range as reported in (Webb, Holyoak, and Lu, 2023) (blue). (2) Even when using the original alphabet, the generative accuracy drops to about 0.3 and below 0.2 for the 'extend sequence' and 'predecessor problems', respectively. (3) The effect of the prompt, which is needed for the synthetic alphabet, has little effect on the generative accuracy. In fact, the performance is the same and even higher for all problem types except 'successor' and 'remove redundant letter'. This indicates that the prompt is having little impact on the performance when introducing the synthetic alphabet.

⁶We further extended our experimentation by exploring variations of the modifications. For instance, when employing an interval of size five and using the real alphabet, GPT-3's performance yielded results similar to those obtained with an interval of size two and the synthetic alphabet (see Figure 4 in the Appendix). Additionally, in an experiment featuring an interval of size one (as in the original paper) and utilizing the synthetic alphabet, performance improved but remained inferior to that observed in the initial paper (see Figure 5 in the Appendix).

4 Discussion

The recent paper, "Emergent analogical reasoning in large language models" (Webb, Holyoak, and Lu, 2023), and subsequent news articles argue that LLMs may have acquired the emergent ability for zero-shot analogical reasoning. We are less certain of these conclusions, given our own follow-up experiments. Our results show low success of GPT-3 in solving letter string problems with simple modifications and with a synthetic alphabet. These results counter the "very general capacity in zero-shot analogical reasoning" of GPT-3 (Webb, Holyoak, and Lu, 2023). If GPT-3 had such a capacity to solve analogy tasks like letter string problems, we would expect to see similar performance for simple extensions of the analogy problems and a modified alphabet.

GPT-3 achieved similar generative accuracy⁷ to the real alphabet in only two out of six problem types ('remove redundant letter' and 'sort'). For these problems, GPT-3 does not need to generate a letter from the full alphabet but only to remove the duplicate letter or to rearrange given letters, which may explain the higher performance. The results on these two task also show that GPT-3 is able to process synthetic alphabets, which is important when interpreting the lower performance for the other letter string problems.

So what explains the high success of GPT-3 in solving the problems on the real alphabet (as used in the original paper) but failure with the synthetic alphabet and with the modified interval size for most of the letter string problems?

Our results suggest that the answer may reside in the training data. GPT-3 performs well for simple analogy problems with the standard English alphabet, which are likely to be present in the training data. However, analyzing the training data and verifying its impact on performance is nearly impossible. Most researchers don't have access to the training data for GPT-3 and, even if researchers had access, it is difficult to rule out the possibility of the examples or derivations residing in the training data. To strengthen claims of human-like reasoning such as zero-shot reasoning (Webb, Holyoak, and Lu, 2023), it is important that the field develop approaches that preclude data memorization.

Zero-shot reasoning is an extraordinary claim that requires extraordinary evidence. We don't see that evidence in our experiments. In addition to comparing GPT-3's performance to human subjects for various analogical problems, one would need to show that the problem or even similar problems do not exist in the training data as noted above. Since this is not realizable today, analogical problem sets that are novel to GPT-3 in both its specificity (specific examples) and its characteristics (pattern of examples) is needed. This does not apply to the letter string problems used in the original paper, as the authors themselves note in the peer review file⁸: "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." To address this desired zero-shot condition, we extend the original letter string tasks to include a synthetic alphabet—one that GPT-3 has

⁷Interestingly, GPT-3 seems to perform better than ChatGPT (August 3 Version). More tests need to be conducted to determine if this holds for all examples.

⁸<https://www.nature.com/articles/s41562-023-01659-w#peer-review>

likely not seen or at least seen as often. GPT-3’s poor performance on these synthetic alphabets is potential evidence against the claimed zero-shot reasoning capacity.

In the peer review file, the authors further note that they ask GPT-3 about these problems as a way of testing their existence in the training data. It makes sense to at least try this, but we find this to be weak evidence, given the large number of possible answers to this question ⁹.

Webb, Holyoak, and Lu (2023) find that GPT-3 exhibits human-like characteristics in analogical reasoning, i.e., decreasing performance with increasing problem complexity. Based on this result, the authors propose that GPT-3 may have developed mechanisms similar to those underlying human intelligence. While this is one possible interpretation, it is important to note that the observation of similar performance between GPT-3 and humans does not inherently validate GPT-3’s human-like, zero-shot analogical reasoning. An alternative explanation could be that the training data contains a scarcity of solutions to complex problems, possibly reflecting the challenges humans encounter with such problems. More work is needed to investigate these various possibilities.

5 Conclusion

Based on their results, (Webb, Holyoak, and Lu, 2023) argue that “large language models such as GPT-3 have acquired an emergent ability to find zero-shot solutions to a broad range of analogy problems.” With full respect to the authors and their work investigating interesting questions, we disagree with this conclusion. We find that GPT-3 fails at solving even the simplest variations of letter string problems. Based on our results, we do not see strong evidence of this emergent ability. Given the current hype surrounding LLMs, we hope this can be used to spur further tests and evaluations of what LLMs can and cannot do.

References

- Mitchell, Melanie (Jan. 2023). *On analogy-making in large language models*. URL: <https://aiguide.substack.com/p/on-analogy-making-in-large-language> (visited on 08/09/2023).
- Webb, Taylor, Keith J. Holyoak, and Hongjing Lu (July 2023). “Emergent analogical reasoning in large language models”. en. In: *Nature Human Behaviour*. ISSN: 2397-3374. DOI: 10.1038/s41562-023-01659-w. URL: <https://www.nature.com/articles/s41562-023-01659-w>.

⁹On the contrary, ChatGPT is able to provide examples of letter string problems, suggesting that it may have seen such examples during training. We include our question and ChatGPT’s answer in the appendix. Important to note is that ChatGPT was trained on more data than GPT-3 so this result only provides circumstantial evidence.

6 Appendix

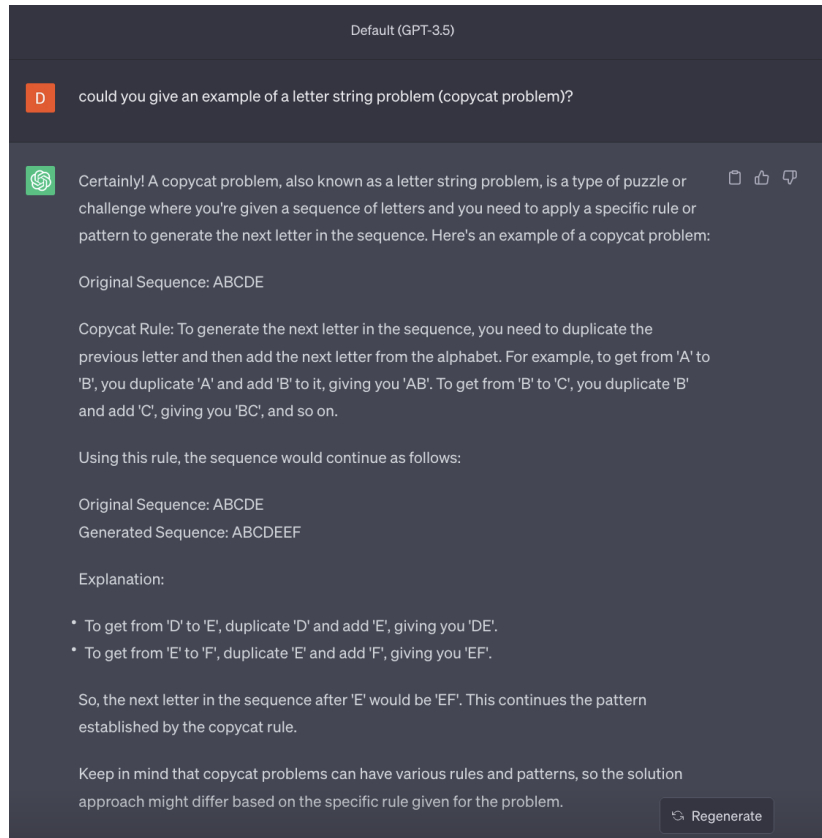


Figure 3: ChatGPT's answer to our question: "could you give an example of a letter string problem (copycat problem)?"

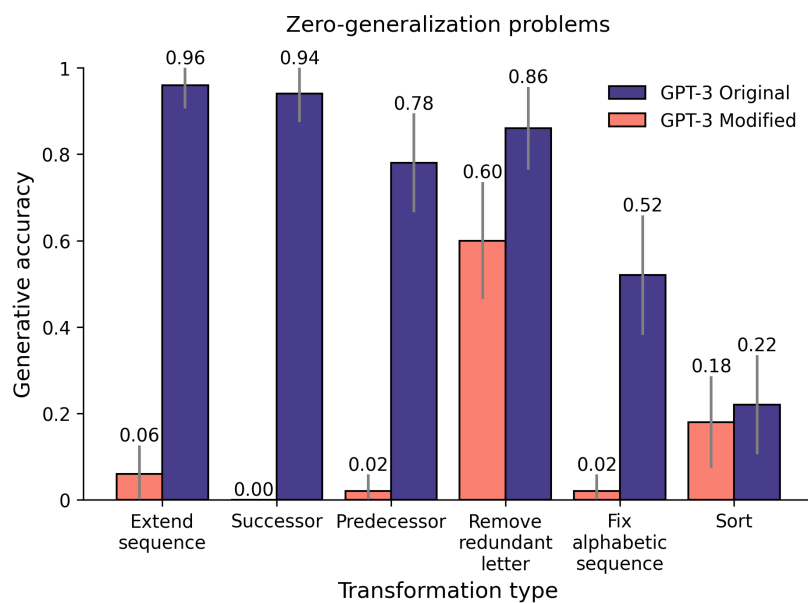


Figure 4: GPT-3 performance on the original (blue) and larger interval of size five (red) zero generalization letter string problems, as a function of transformation type. In both settings, we used the real alphabet without prompt modification. The figure and the order of the transformation types correspond to Figure 6a in the original paper. These results reflect an average performance for $N=50$ instances. For the modified problems, only 'remove redundant letter' achieves an accuracy greater than 0.1.

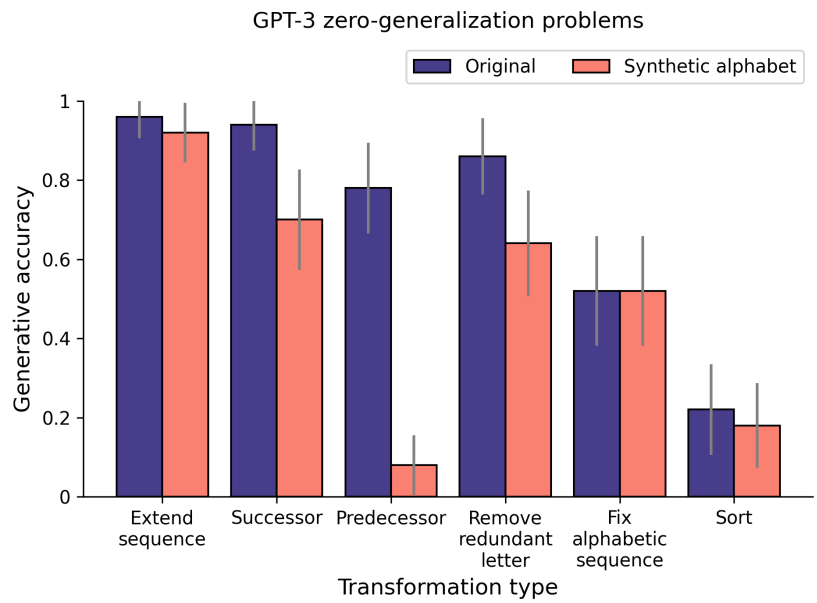


Figure 5: GPT-3 performance for zero generalization letter string problems on the real (blue) and the synthetic alphabet (red) without modification of the interval size. The figure and the order of the transformation types correspond to Figure 6a in the original paper. These results reflect an average performance for N=50 instances.