Probing Large Language Models to Perform Analogical Transformation

François Olivier[1,*], Miguel Couceiro[2] and Zied Bouraoui[1]

[1]CRIL CNRS & Univ Artois, France
[2]U.Lorraine, CNRS, LORIA & IST, U.Lisbon

Abstract
In this paper, we investigate the capability of LLMs to perform simple analogical reasoning on a manually constructed mathematical question-answering dataset. This dataset contains pairs of numbers constructed following a mathematical pattern that can be easily found using analogy based inference. We test GPT-4 through ChatGPT, and show that it needs no more than three pairs of numbers to discover the underlying pattern, except when that pattern involves the application of more than two operations. For the latter instances, GPT-4 tends to engage in increasingly complex mathematical reasoning (*e.g.,* using square roots, working with reals) instead of exploiting simpler mathematical expressions that lead to the solution.

Keywords
Large Language Model, Mathematical Reasoning, Reasoning with Analogy

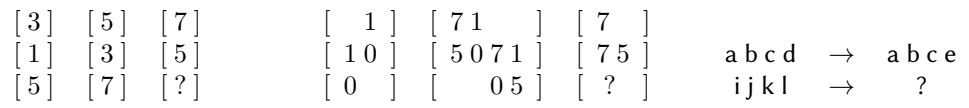
1. Introduction

Analogical reasoning is an exceptional cognitive ability that allows humans to solve complex problems by drawing upon knowledge from one domain and applying it to a different, yet analogous, domain [1, 2]. This process involves identifying relevant similarities while ignoring differences between the two domains in order to make inferences and draw conclusions based on these similarities. The domains considered as well as the complexity of the knowledge can vary significantly in analogical tasks (*e.g.,* ranging from simple words to entire stories for linguistic analogies).

Within the field of artificial intelligence, word embedding models represented a major step forward for analogical reasoning, as they significantly improved the performance on word analogy tasks of the form $a : b :: c : d$ (*e.g.,* tarantula:spider::bee:?) [3]. The correctness of these inferences, however, has been shown to decrease with the variety of relation types considered [4, 5]. With the advent of Large Language Models (LLMs), the study of high-level cognitive abilities became the subject of increasing research in various fields [6, 7]. In this respect, LLMs are tested on their capacities to make analogies between words, sentences and short stories, but also on their logico-mathematical capability to solve small puzzles and codes such as Raven's IQ tests [8].

IARML@IJCAI'2024: Workshop on the Interactions between Analogical Reasoning and Machine Learning, at IJCAI'2024, August, 2024, Jeju, South Korea
*Corresponding author.
olivier@cril.fr (F. Olivier); miguel.couceiro@loria.fr (M. Couceiro); zied.bouraoui@cril.fr (Z. Bouraoui)

For instance, [9] test zero shot reasoning with GPT-3 for matrix problems (Figure 1(a) and (b)), letter string problems (c), word analogy tasks as described above, analogies between stories, and analogical problem-solving (*i.e.* trying to solve a problem by first hearing a story that is analogical to the solution of the problem).



(a) Digit transformation problem with a progression rule.

(b) Logic problem based on a OR rule.

(c) Letter string analogy problem.

Figure 1: Each problem is read row by row and the goal is to find the element that replaces the '?' symbol in the last row. The pattern to find in (a) is a progression by two. In (b), the digits in the middle column are defined as the union of the sets present in the other columns. In (c), the last letter of the sequence is transformed into the successor letter in the alphabet.

As detailed in [9], GPT-3 showed impressive results that even outperform human results in several logical tasks, but keeps struggling with analogical reasoning at the level of stories in natural language [10, 9]. These observations bring about fundamental questions about whether a genuine analogical reasoning ability emerges in LLMs or if they simply become better at analogical tasks by merely improving their capacities to exploit heuristic short-cuts [11, 12]. Research such as [13] enhanced these critics by showing some discrepancies between the resolution of word analogy tests and relational structure identification tests, therefore indicating that LLMs do indeed overlook the underlying structures while solving analogical tasks. In order to test these models in a more consistent way, several benchmarks such as the **Sc**ientific **A**nalogical **R**easoning with structure abduction (SCAR) [13], or the STORYANALOGY corpus [10], have been proposed. With the SCAR benchmark, authors intend to place LLMs on a par with humans by forcing them to abstract the structure underlying different analogical domains. The STORYANALOGY corpus extends the Structure Mapping Theory of Gentner [1] to establish a clear and specific way to evaluate story analogies on longer texts. These more stringent approaches reveal poor performances of LLMs, often outperformed by humans. Although some prompting techniques enable LLMs to improve their analogical reasoning capacity [14], it remains a challenge for them to reach the level of abstraction required to fully cover this fundamental cognitive ability.

In this work, we further probe the analogical reasoning abilities of an LLM by testing it on mathematical problems in a few shot approach. Besides extending our assessment of LLMs for analogical reasoning involving mathematics, our tests also provide insights into the inner workings of these models for solving arithmetic analogical equations.

2. Analogical transformation with LLMs

We consider analogical questions that involve a series of pairs (A, B) presented in an incremental way. Given a pair (A_i, B_i), the goal for the LLM is to find the pair (A_{i+1}, B_{i+1}) that maintains the same relationship as the previously seen pairs, and to explain this relationship if necessary.

We formulate the analogical question as a simple 1-step analogical transformation. The following patterns are used to generate examples, where t is an integer and a and b are constants:

1. $t : t + a$
2. $t + a : t \times a$
3. $t \times a : t + a$
4. $t : t^2$
5. $t + a : t^2$
6. $t + a : t^a$
7. $t \times a : t^2 + a$
8. $t + b : (t \times a) + b$
9. $t \times b : t^a + b$
10. $t \times a : (t \times a) + (t \times b)$

We make use of GPT-4 through the ChatGPT interface and we employ the following steps to conduct our tests. We first provide the following instruction:

"I will present a series of analogies that are pairs of the form A:B, C:D, E:F, etc. Your goal is to find the mathematical logic behind these pairs so that you can predict the next pair of the list."

The method used is then to provide a first pair with $t = 2$. If the correct relation for $t = n$ is not found or the next pair is only found by chance, the pair constructed with $t = n + 1$ is added to the pairs given.

Example 1. *For the analogical transformation $t + a : t \times a$ (for a constant $a = 3$), we first provide the pair $5 : 6$. The relation guessed is $B = A + 1$, which is not the relation expected. Providing the following pair $6 : 9$ does not yield the expected relation either. Finally, providing $7 : 12$ enables GPT-4 to find the relation $B = 3A - 9$, which is equivalent to $t + a : t \times a$, and consequently the correct next pair $8 : 15$ is predicted.*

For the analogical transformations 1-6, 8 and 10, only two or three pairs are necessary for GPT-4 to find a correct relation. For instance, after the two pairs $2 : 5$, $3 : 6$ (Transformation 1 with $a = 3$), the model predicts the relation $B = A + 3$, or after the two pairs $4 : 8$, $5 : 11$ (Transformation 8 with $a = 3$ and $b = 2$), it finds the relation $B = 3A - 4$. Note that all of its answers are given in the form $B = f(A)$, where $f(A)$ represents a function applied on A to obtain B. This reasoning tends to show that GPT-4 treats the elements of pairs as constants, and consequently, does not easily decompose them into smaller numbers. This observation is confirmed with transformations 7 and 9, for which GPT-4 significantly increases the complexity of the functions tested (*e.g.,* using square roots or divisions that result in decimals numbers) without being able to find a coherent solution with simple decomposition of numbers, even after the fifth pair. For these transformations, a function involving more than two operations had to be derived (*e.g.,* for the pairs $6 : 7$, $9 : 12$, $12 : 19$, etc (Transformation 7 with $a = 3$), the function is $B = ((A/3)^2) + 3$).

3. Conclusion

In this preliminary work, we further examined the capability of LLMs to perform analogical transformations in order to answer mathematical analogical questions. Our tests indicate that GPT-4 shows good performance for analogical transformations when they do not exceed a certain level of complexity, but lacks the ability to exploit different mathematical expressions when the analogical transformations are less direct. In future research, we plan to investigate a wider range of large language models and delve into further analogical tasks.

Acknowledgments

This work was partially supported by the ANR projects ANR-22-CE23-0002 (ERIANA) and ANR-22-CE23-0023 (AT2TA).

References

[1] D. Gentner, Structure-mapping: A theoretical framework for analogy, Cognitive science 7 (1983) 155–170.

[2] D. R. Hofstadter, E. Sander, Surfaces and essences: Analogy as the fuel and fire of thinking, Basic books, 2013.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).

[4] A. Gladkova, A. Drozd, S. Matsuoka, Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't., in: Proceedings of the NAACL Student Research Workshop, 2016, pp. 8–15.

[5] T. Czinczoll, H. Yannakoudakis, P. Mishra, E. Shutova, Scientific and creative analogies in pretrained language models, arXiv preprint arXiv:2211.15268 (2022).

[6] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, CoRR abs/2307.09288 (2023). URL: https://doi.org/10.48550/arXiv.2307.09288. doi:10.48550/ARXIV.2307.09288. arXiv:2307.09288.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever,

D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[8] W. C. Beltran, H. Prade, G. Richard, Constructive solving of raven's IQ tests with analogical proportions, Int. J. Intell. Syst. 31 (2016) 1072–1103.

[9] T. Webb, K. J. Holyoak, H. Lu, Emergent analogical reasoning in large language models, Nature Human Behaviour 7 (2023) 1526–1541.

[10] C. Jiayang, L. Qiu, T. H. Chan, T. Fang, W. Wang, C. Chan, D. Ru, Q. Guo, H. Zhang, Y. Song, et al., Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding, arXiv preprint arXiv:2310.12874 (2023).

[11] M. R. Petersen, L. van der Plas, Can language models learn analogical reasoning? investigating training objectives and comparisons to human performance, arXiv preprint arXiv:2310.05597 (2023).

[12] C. E. Stevenson, M. ter Veen, R. Choenni, H. L. van der Maas, E. Shutova, Do large language models solve verbal analogies like children do?, arXiv preprint arXiv:2310.20384 (2023).

[13] S. Yuan, J. Chen, X. Ge, Y. Xiao, D. Yang, Beneath surface similarity: Large language models make reasonable scientific analogies after structure abduction, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 2446–2460. URL: https://doi.org/10.18653/v1/2023.findings-emnlp.160. doi:10.18653/V1/2023.FINDINGS-EMNLP.160.

[14] M. Yasunaga, X. Chen, Y. Li, P. Pasupat, J. Leskovec, P. Liang, E. H. Chi, D. Zhou, Large language models as analogical reasoners, arXiv preprint arXiv:2310.01714 (2023).