Thesis for Bachelor's Degree

MC-LARC Benchmark to Measure LLM Reasoning Capability

Shin, Donghyeon (신 동현 申 棟賢)

School of Electrical Engineering and Computer Science

Gwangju Institute of Science and Technology

2023

MC-LARC Benchmark to Measure LLM Reasoning Capability

거대언어모델의 추론능력 평가를 위한 MC-LARC 데이터셋

MC-LARC Benchmark to Measure LLM Reasoning Capability

Advisor : Professor Kim, Sundong

by

Shin, Donghyeon

School of Electrical Engineering and Computer Science
Gwangju Institute of Science and Technology

A thesis submitted to the faculty of Gwangju Institute of Science and Technology in partial fulfillment of the requirements for the degree of Bachelor of Science in Electrical Engineering and Computer Science

Gwangju, Republic of Korea

2023. 12. 22.

Approved by

Professor Kim, Sundong

Committee Chair

MC-LARC Benchmark to Measure LLM Reasoning Capability

Shin, Donghyeon

Accepted in partial fulfillment of requirements
for the degree of Bachelor of Science

December. 22. 2023.

Committee Chair ______________________
Prof. Sundong Kim

Committee Member ______________________
Prof. Jeany Son

Contents

Chapter 1. Introduction

At the present moment, large language models have demonstrated high performance in various natural language processing domains. However, there have been criticisms regarding the inferential capabilities of these models [1]. Some prior research has evaluated the lack of inferential ability in large language models based on their poor performance on the Abstraction and Reasoning Corpus (ARC) dataset [2], which has been proposed as a benchmark for evaluating reasoning skills [3] [4][5]. However, it is important to note that the process of evaluating the inferential ability of large language models in tasks involves two stages: 1) inferring relationships between given data, and 2) generating sentences to express those inferences. Therefore, ARC dataset may not serve as appropriate benchmarks for assessing inferential abilities since they encompass both stages. Hence, in this paper, we aim to propose a new dataset called MC-LARC (Multiple-Choice LARC) that allows for a more appropriate evaluation of inferential abilities in large language models by focusing solely on the inferential task, excluding the sentence generation phase.

Our MC-LARC dataset is based on ARC. The original ARC consists of small images represented as two-dimensional matrices, and it presents a problem of inferring rules between input images and output images. LARC (Language-complete ARC) [6] extended ARC by describing the input images and rules in sentences, thus bridging the gap between ARC and natural language processing. However, the existing LARC dataset suffered from the issue of being poorly refined and not containing sufficient information about the problems, as it was collected in an uncontrolled environment through Amazon Mechanical Turk crowdsourcing. To address this issue, we manually refined the existing LARC dataset and then used this refined dataset along with the ChatGPT4-32k model to create the MC-LARC dataset.

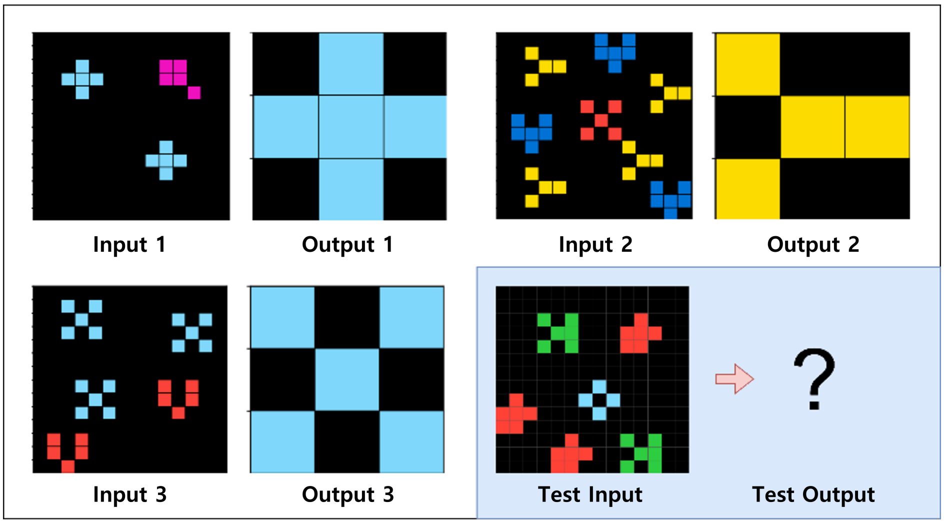


Figure 1.1: The goal is to infer the correct output image (on a blue background) by examining the input and output images and deducing the underlying rule (on a white background).

Chapter 2. Dataset

2.1 Original ARC Dataset

The Abstraction and Reasoning Corpus (ARC) dataset [2] was created for the purpose of measuring the intelligence of computer systems. This dataset demands deep thinking and inference based on complex prior knowledge such as mathematical abilities, geometric understanding, and topological comprehension. Figure 1.1 provides an example of ARC data, where the goal is to derive a common rule from three examples and apply it to infer the correct output image for a given test input image. Each problem includes 2 to 5 pairs of input and output images as examples. The original ARC dataset consists of 400 training data, 400 evaluation data, and 200 test data. ARC dataset is represented as 2-dimensional matrices, and when visualized, it looks like the image shown in Figure 1. The 400 training data are composed of different problems.

In creating the descriptive sentences for the input images in MC-LARC, which is proposed in this research, we referred to the visualized original ARC input images as seen in Figure 1.

2.2 Text Data - LARC Dataset

The LARC dataset consists of descriptions for each of the 400 training data from the original ARC dataset, including 1) descriptions of the input images and 2) descriptions of the rules between the input and output images. Additionally, a confidence rating item was added to indicate how confident the participants were in the sentences they provided.

However, the original LARC dataset has limitations in providing insufficient information for problem-solving. Also, the original LARC dataset was created by multiple non-experts, leading to inconsistencies. For instance, there were cases where different expressions were used for the same color pixels. Furthermore, despite high confidence ratings, there were instances of irrelevant text being included in the descriptions, as seen in the left description of Figure 3.1, making the data unreliable even when considering the confidence ratings. Therefore, the problem description section of LARC was refined to contain meaningful information.

Chapter 3. Multiple Choice LARC (MC-LARC) Dataset

The proposed MC-LARC is a benchmark that resembles LARC in describing ARC problems but in the form of multiple-choice questions, including incorrect answer choices. Similar to the original LARC, it includes descriptions of input images and explanations of problem-solving rules. To address the issues mentioned earlier with the original LARC, we first conducted a primary filtering of the descriptions of input images and rule sentences using LARC's confidence data. Then, we further refined them through human-level verification. For the rule sentences, we constructed multiple-choice questions consisting of one correct answer and four incorrect answers. The goal was to evaluate the inferential ability of large language models by solving these multiple-choice questions, skipping the text generation process.

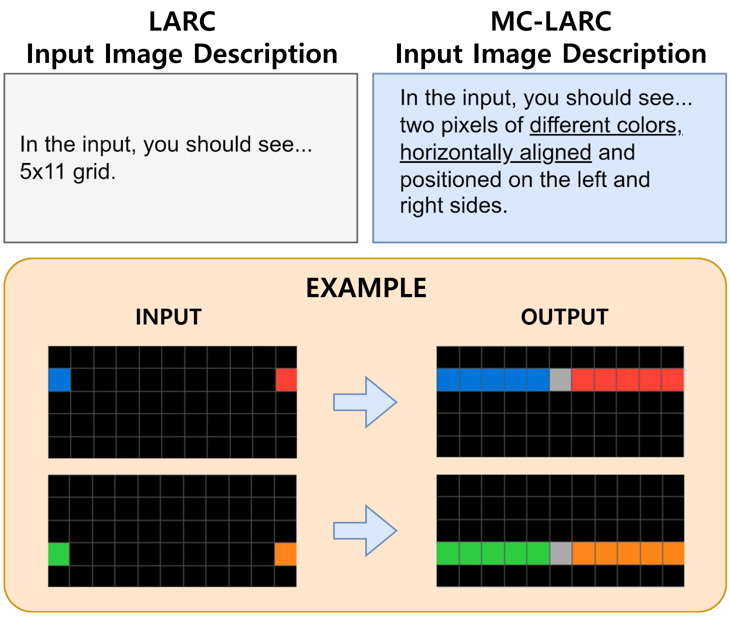


Figure 3.1: Based on the input image of the example problem below, we have modified the input image description from the original LARC (top left) to the MC-LARC (top right).

3.1 Refining Sentences Describing Input Images

To create descriptions for the input images in MC-LARC, we conducted a refinement process for the sentences describing input images from the existing 400 LARC instances. In order to rigorously evaluate the model's inferential abilities, we ensured that the descriptions for input images did not imply the rules, and we refined the sentences based on five criteria: color, object information, numerical details, geometry and topology, and common sense. Here is a detailed explanation of each criterion:

1. **Color**: Considering that many problems in the ARC dataset are related to rules involving colors, recognizing information about the colors of pixels is important. We included color information accordingly.

2. **Object Information**: A significant portion of ARC problems is related to objects, with object information being crucial for problem-solving in about 50% of the cases [3]. Objects can be defined as sets of adjacent pixels, pixels of the same color, or similar patterns.

3. **Numerical Details**: Some problems require the recognition of pixel or object counts to be solvable. Therefore, we included information about the number of pixels or objects to facilitate the recognition of such details.

4. **Geometry and Topology**: For problems where the arrangement of pixels forms geometric shapes like triangles and rectangles or involves spatial relationships between pixels, we described this information.

5. **Common sense**: Information related to everyday physics concepts or patterns such as radial patterns, checkered patterns, etc., was written based on common human knowledge.

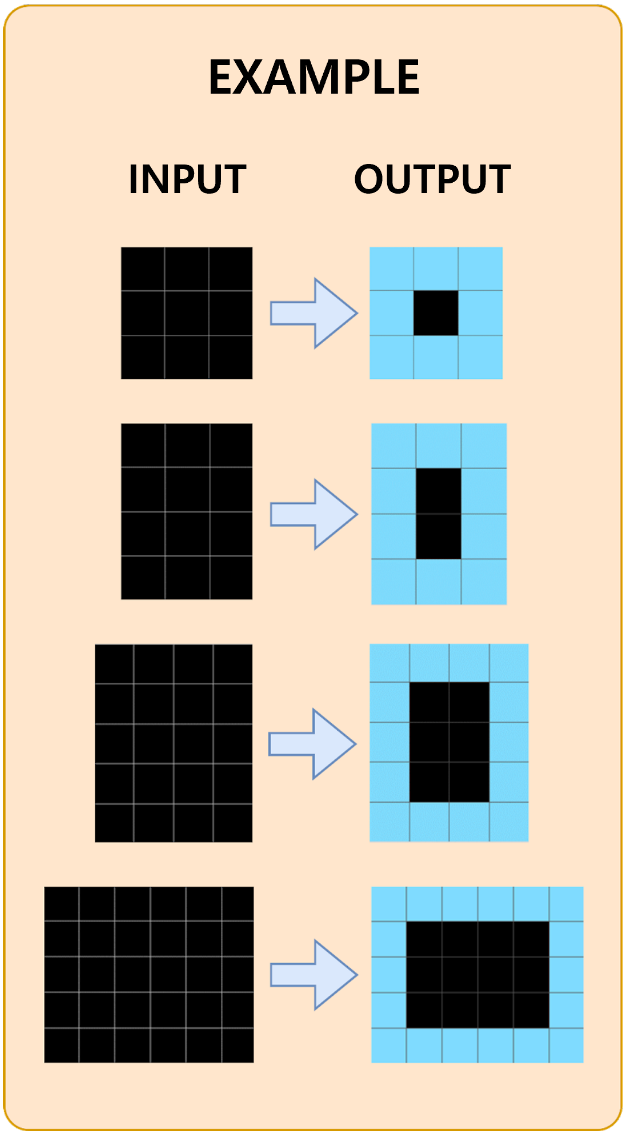
3.2 Sentence Augmentation for Rules (Explanations)

To utilize the large language model ChatGPT4-32k, we augmented sentences (answers) in the refined LARC with four additional distractor sentences that are similar in structure but completely different in meaning, as shown in Figure 3.2.

In generating these distractors, our research imposed two major constraints at the prompt level. Firstly, we prevented the creation of distractors by merely replacing words with synonyms, thus avoiding generating incorrect answers that are simply synonymous. Secondly, we provided background knowledge about the ARC dataset environment to minimize the generation of sentences that do not fit the context of the ARC dataset. Through the prompts described above, we augmented distractor sentences that are similar to the correct answers but have entirely different meanings.

Additionally, to prevent the large language model from identifying correct answers by recognizing specific parts only present in the correct answers, we standardized the sentence format. For example, if the correct answer sentence is described as "To make the output, you have to..." and another distractor is written as "To transform into an output image, one must", it's possible to solve the problem by just looking for the sentence "To transform into an output image, one must", without making the necessary inferences. Similarly, we randomized the order of the correct answers. If one could guess the correct answer based on its order, it would be difficult to say that the model actually made an inference.

MC-LARC SOLUTION



To make the output, you have to... create a light blue border around the grid. So, any squares that are touching the edges are filled in with light blue. The rest of the squares inside the light blue border remain black.

To make the output, you have to... fill all squares with black and then create a border in yellow. The space inside the yellow border remains black.

To make the output, you have to... create a red grid in the center, leaving a black border around the edges. The rest of the squares inside the red grid remain black.

To make the output, you have to... leave all squares black without creating any form of border.

To make the output, you have to... create a pattern within the grid using different colors. The edge color does not differ from the inner ones.

Figure 3.2: You can see the MC-LARC with one correct answer choice (Green background) and four incorrect answer choices (Red background) for the example question.

Chapter 4. Experiment and Results

In this research, we measured how well humans and a large language model (referred to as MC-LARC) solve Multiple Choice Logical Reasoning Comprehension (MC-LARC) questions. We compared the difference in performance when providing hints about the input images and when not providing any hints. Through this, we evaluated the quality of our MC-LARC by comparing it with LARC.

4.1 Evaluation of MC-LARC - ChatGPT

I conducted language model experiments using GPT-4 provided by ChatGPT Plus. Currently, ChatGPT Plus allows for attaching up to 4 images when asking questions, so I provided a maximum of 4 input image-output image pairs.

The experiments were conducted in two ways, one with modified input image description data and background knowledge explaining the ARC dataset, and the other without such background knowledge.

Table 4.1: The (Image + Input description) and (Image) show the accuracy rates in solving questions that involve selecting solutions for the ARC by providing the MC-LARC dataset. The (Original ARC data) shows the accuracy rates when directly modifying the ARC's 2D matrices.

	Image + Input description	**Image**	**Original ARC data**
ChatGPT4	337/400 **(84.25%)**	316/400 **(79.00%)**	-
GPT4-0613	-	-	77/800 **(9.625%)**

Looking at the Table 4.1, ChatGPT-4 achieved approximately 80% accuracy in both cases. This demonstrates a very high success rate when compared to prior research [5] that attempted to solve the original ARC using large language models.

4.2 Evaluation of MC-LARC - Human

MC-LARC targeting humans showed an accuracy rate of 90.75%, which was higher than the 83.8% accuracy rate reported in prior research [7] on the ARC dataset.

Table 4.2: This is the result of the MC-LARC dataset experiment conducted on humans. The difficulty levels were marked from 1 to 5, and the ratio of **(# of correctly answered / # of questions)** is presented.

Difficulty	**1**	**2**	**3**	**4**	**5**
	149/155 **(96.1%)**	91/102 **(89.2%)**	67/72 **(93%)**	38/41 **(92.7%)**	18/30 **(60%)**

Table 4.2 displays the distribution of difficulty ratings assigned by humans and the number of correctly answered questions for a total of 400 questions.

Chapter 5. Future Research Proposals

Firstly, we propose research on a model capable of generating descriptions for input images at a human-level quality. During the refinement of the MC-LARC dataset, the cost was significant because we manually reviewed and crafted selected sentences. Additionally, for evaluating performance in the ARCathon [8], the test dataset lacks textual descriptions for input images, necessitating the need for a model that can generate textual data directly. Therefore, we aim to implement an image captioning model to overcome the aforementioned limitations and establish a foundation for transforming image inference problems into text inference problems.

Finally, we intend to explore the use of a multimodal model structure that utilizes information from both image and text data in solving ARC problems. Leveraging image information can enhance the reasoning capabilities of large language models, and furthermore, by utilizing the parameters of this model, we aim to build a model capable of generating images, ultimately aiming to solve ARC problems effectively.

Chapter 6. Conclusions

In this research, we transformed ARC problems from image inference tasks into text inference tasks. Additionally, taking into account the limitations of existing large language models in terms of their inference capabilities, we proposed the MC-LARC multiple-choice question dataset, excluding sentence generation tasks and focusing solely on evaluating inference abilities. Recognizing the potential limitations in the reliability of the LARC dataset used for creating MC-LARC, we are aware of the need for human-level evaluation and refinement of the dataset.

Nevertheless, through various efforts like MC-LARC, we aim to contribute to the exploration of the strengths, weaknesses, and limitations of large language models as artificial general intelligence (AGI). We anticipate that these contributions will further the development of artificial general intelligence.

Summary

MC-LARC Benchmark to Measure LLM Reasoning Capability

거대언어모델(LLM)은 현재 가장 주목 받고 있는 인공지능 모델입니다. 하지만 거대언어모델이 실제 인간 수준의 추론 능력을 가지고 있는지는 아직 밝혀지지 않았으며, 다양한 연구가 진행되고 있습니다. 따라서 거대언어모델이 인간 수준의 추론 능력을 가지고 있는지 평가하여 인공일반지능(AGI)로의 가능성을 보이는 것은 중요하다고 할 수 있습니다. 이를 위해 본 연구에서는 거대언어모델의 추론 능력을 적합하게 평가하기 위한 MC-LARC 데이터셋을 구축하였습니다.

MC-LARC 데이터셋은 기존에 인공지능의 지능을 측정하기 위해 만들어진 ARC 데이터셋을 기반으로 만들어진 데이터셋입니다. 하지만 기존의 ARC 데이터셋은 이미지 기반 데이터셋으로, 자연어에 특화된 거대언어모델에 직접 활용하기 적합한 형태가 아닙니다. 또한 ARC 데이터셋 문제를 해결하기 위해서는 적합한 형태의 정답을 생성해야하는 어려움이 있습니다. 이러한 한계를 극복하기 위하여 MC-LARC 데이터셋은 자연어로 구성된 데이터셋으로 만들었으며, 생성 문제에서 선택 문제로 변형 및 확장하였습니다.

MC-LARC 데이터셋은 자연어를 포함한 데이터셋입니다. 거대언어모델의 추론 능력을 평가하는 용도 뿐만 아니라, Multi-modal 연구, 거대언어모델 연구에 적극적으로 활용할 수 있을 것으로 기대하고 있습니다. 궁극적으로 인간 수준의 추론 능력을 지닌 인공지능 모델로 나아가 사람 처럼 생각할 수 있는 인공일반지능에 도달할 수 있기를 기대하고 있습니다.

감 사 의 글

논문 작성을 처음 접해서 방황하고 헤매던 제게 조언을 아끼지 않으신 김선동 지도교수님께 감사드립니다. 또한 귀한 시간을 내어 논문 심사를 봐주신 손진희 교수님께 감사드립니다. 이 외에도 함께 연구한 황산하, 이석기, 김윤호, 이승필 그리고 제게 도움을 주신 연구실 동료 모든 분들께 감사드립니다.

약 력

이　　　름: 신동현

생 년 월 일: 1997년 4월 11일

출　생　지: 전라북도

주　　　소: 경기도 평택시 추담로 58-78

학　　력

2014. 3. – 2016. 2.　평택고등학교

2018. 2. – 2024. 2.　광주과학기술원 전기전자컴퓨터공학부 (학사)

References

1. Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and Fate: Limits of Transformers on Compositionality. 2023.

2. François Chollet. On the Measure of Intelligence. *arXiv:1911.01547*, 2019.

3. Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias B Khalil. LLMs and the Abstraction and Reasoning Corpus: Successes, Failures, and the Importance of Object-based Representations. *arXiv:2305.18354*, 2023.

4. Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The ConceptARC benchmark: Evaluating Understanding and Generalization in the ARC Domain. *arXiv:2305.07141*, 2023.

5. Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. *arXiv:2307.04721*, 2023.

6. Samuel Acquaviva, Yewen Pu, Marta Kryven, Theodoros Sechopoulos, Catherine Wong, Gabrielle Ecanow, Maxwell Nye, Michael Tessler, and Joshua B. Tenenbaum. Communicating Natural Programs to Humans and Machines. In *NeurIPS*, 2022.

7. Aysja Johnson, Wai Keen Vong, Brenden M. Lake, and Todd M. Gureckis. Fast and Flexible: Human Program Induction in Abstract Reasoning Tasks. In *CogSci*, 2021.

8. Michael Hodel. ARC: Where Do We Stand Today?, 2023.