MC-LARC Benchmark to Measure LLM Reasoning Capability

DongHyeon Shin [1]

[1]Gwangju Institute of Science and Technology

Abstract

The Abstract Reasoning Corpus (ARC) poses a challenging problem in artificial intelligence due to its limited data, making it difficult for conventional AI models that require large datasets. On the other hand, large language models have demonstrated high performance across various domains. Therefore, in this research, we propose a new dataset called MC-LARC to leverage the inferential capabilities of large language models for solving ARC. The MC-LARC dataset introduced in this study consists of 1) one sentence describing an example input image from ARC and 2) five sentences explaining the problem-solving rules. The image description sentences were manually created by humans, while the problem-solving rule sentences were generated using GPT-4 32k. Subsequently, we conducted inference ability tests using MC-LARC to determine whether ChatGPT-4 and humans can select appropriate descriptions for a given image.

Introduction

ARC

The Abstraction and Reasoning Corpus (ARC) dataset 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.

LARC

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. Furthermore, there were instances of irrelevant text being included in the descriptions, as seen in the left description of Figure 1, making the data unreliable. Therefore, the problem description section of LARC was refined to contain meaningful information.

Research objectives

- **Objective 1:** Creating a new dataset, MC-LARC, to evaluate the inferencing ability of the LLM.
- **Objective 2:** Expanding ARC research into the text domain.

MC-LARC Dataset

In order to properly assess the inference capabilities of LLMs, it was necessary to expand inference ability evaluation benchmark datasets to the text domain. Therefore, in this study, we transformed the ARC dataset, which could only be represented in 2D matrices or images, into the MC-LARC dataset, a set of multiple-choice questions in the text domain.
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 refined them through human-level verification. The goal was to evaluate the inferential ability of large language models by solving these multiple-choice questions, skipping the text generation process.

Input Description

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.**

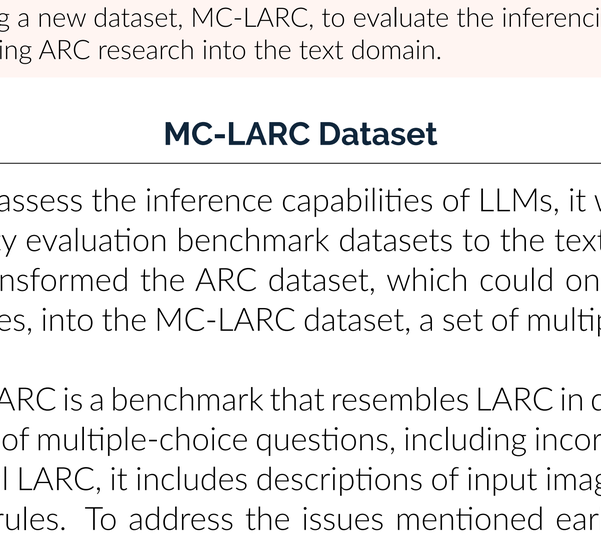

Figure 1. Based on the input image of the example problem below, it was modified from the original LARC input image description (top left) to the MC-LARC input image description (top right).

MC-LARC Options

In generating 4 distractors, our study imposed two major constraints at the prompt level.

- Preventing of **Synonymous** Distractors
- Incorporation of **Background Knowledge** about ARC Dataset

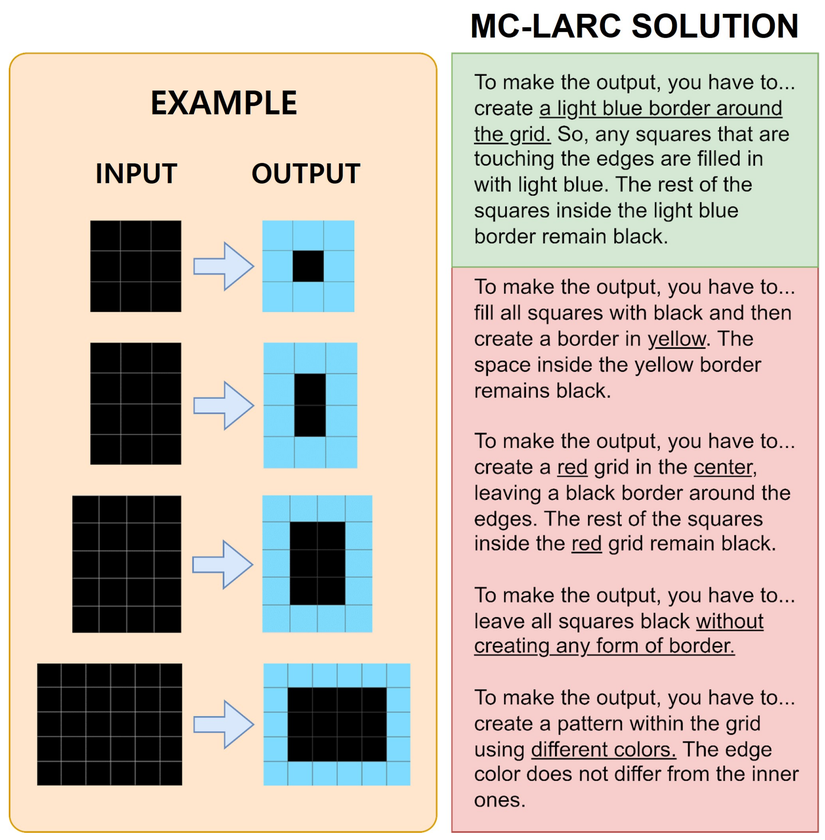

Figure 2. To utilize the large language model ChatGPT4-32k, we augmented sentences (answers) in the refined LARC with four additional distractor sentences that are similar in structure but completely different in meaning, as shown in red part.

Results

Table 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%)**

Table 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%)**	102/91 **(89.2%)**	67/72 **(93%)**	38/41 **(92.7%)**	18/30 **(60%)**

Discussions

- The original ARC problem could not achieve high accuracy using ChatGPT because it had to be modified directly to the ARC dataset. However, the MC-LARC dataset drastically reduced the difficulty of the problem by changing the problem to a multiple-choice problem with the correct answer, and actually showed very high accuracy.
- It is not possible to know from this results whether the LLM actually utilized its reasoning ability and solved the problem through appropriate steps. Therefore, additional research is needed in this regard.
- Looking at Table 1 and Table 2, the MC-LARC dataset is an easy dataset for both LLMs and humans. In the case of ChatGPT4, it showed an accuracy of about 80%, and in the case of humans, it showed an accuracy of about 90%.
- Research on improving the difficulty of the MC-LARC dataset will be needed in the future.

Conclusions

- **Proposal of MC-LARC Dataset focusing on Reasoning Capabilities**: A multiple-choice question dataset, MC-LARC, was proposed to evaluate reasoning capabilities of LLM.
- **Transformation of ARC Problem**: The study transformed the ARC problem from an image inference issue to a text-image inference problem.
- **Reliability Issues with LARC Dataset**: There were concerns about the reliability of the LARC dataset, which was used to create MC-LARC.
- **Contribution to AGI**: MC-LARC are expected to contribute to identifying the strengths, weaknesses, and limitations of large language models in the context of Artificial General Intelligence (AGI).

shindong97411@gmail.com