"How far has the reasoning ability of the Large Language Model (LLM) come?" GIST develops a quantitative evaluation method for LLM reasoning ability

- AI Graduate School Professor Sundong Kim's team develops LLM reasoning ability evaluation framework based on Language of Thought Hypothesis (LoTH)... LLM has some reasoning ability, but still shows significant limitations compared to humans

- Analysis of AI's logical thinking and problem-solving process... Expected to contribute to the development of human-level reasoning ability - Published in international academic journal 《ACM Transactions on Intelligent Systems and Technology》

▲ (Counterclockwise from the left in the front row) Professor Sundong Kim, student Seungpil Lee, student Donghyeon Shin, and researcher Sejin Kim

To what extent can artificial intelligence imitate human reasoning ability*? OpenAI's GPT-4, a large language model (LLM)* applied to ChatGPT, has made great progress in language ability and memory, but is still evaluated as having limited actual logical thinking and reasoning ability.

In particular, the definition of LLM's reasoning ability is ambiguous, and existing evaluation methods are mainly result-oriented, so it is not clear how to objectively and comprehensively evaluate how LLM thinks and reasons.

* reasoning ability: One of the important factors in evaluating the performance of a large-scale language model, this evaluates the ability of the model to derive logical conclusions based on given information, solve problems, and generate answers to complex questions.

* large language model (LLM): Refers to generative language models using large parameters such as ChatGPT and Claude.

The Gwangju Institute of Science and Technology (GIST, President Kichul Lim) announced that Professor Sundong Kim's research team in the AI Graduate School has developed a new framework that can quantitatively measure the reasoning ability of LLM. (https://llm-on-arc.pages.dev/).

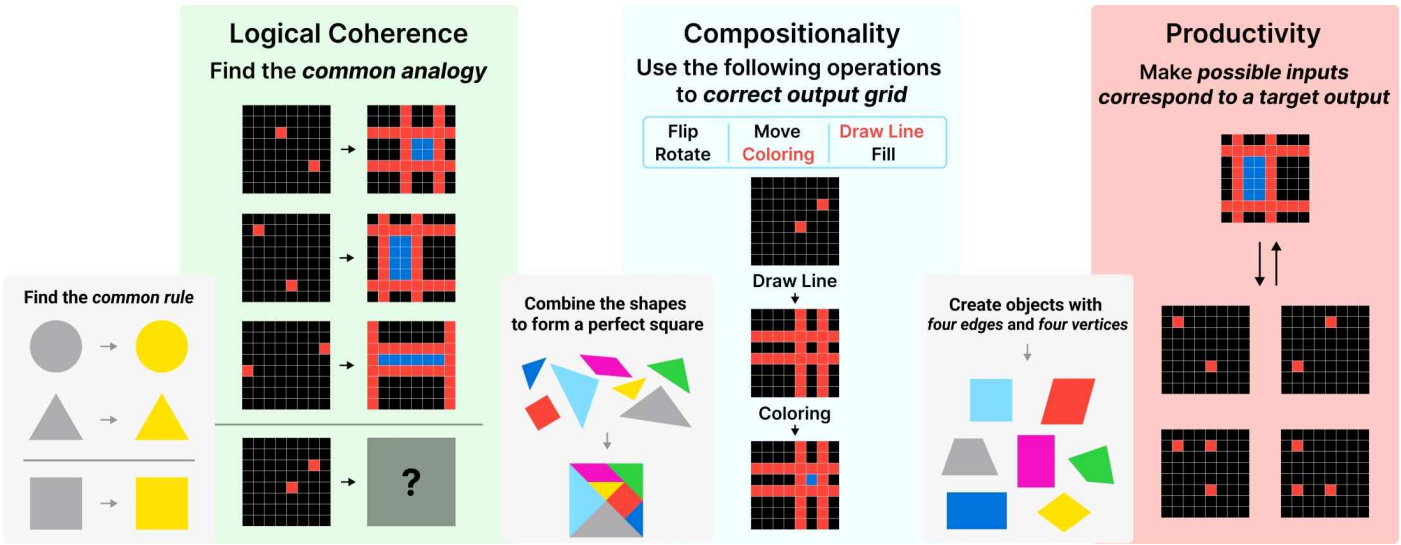
The research team proposed a method to evaluate LLM's reasoning process based on the 'Language of Thought Hypothesis (LoTH)' in cognitive psychology, which states that human cognitive processes are mediated by 'thought language'.

According to this hypothesis, human reasoning processes have three characteristics*: ▲ logical consistency ▲ composition ▲ generativeness. Focusing on these three elements, the research team derived a new approach to evaluate LLM's reasoning and contextual understanding abilities in a process-centered manner using the benchmark* dataset ARC*.

* ▲ Logical consistency refers to the ability to maintain logical consistency in the reasoning process and results, ▲ composition refers to the ability to construct complex ideas by combining simple elements, and ▲ generativeness refers to the ability to infinitely generate new expressions that are not visible in observed data.

* benchmark: This refers to a standard dataset that evaluates the performance of LLM in the field of generative artificial intelligence (AI), and the benchmark score is a numerical value that evaluates how close a specific LLM is to producing the correct answer for the benchmark dataset.

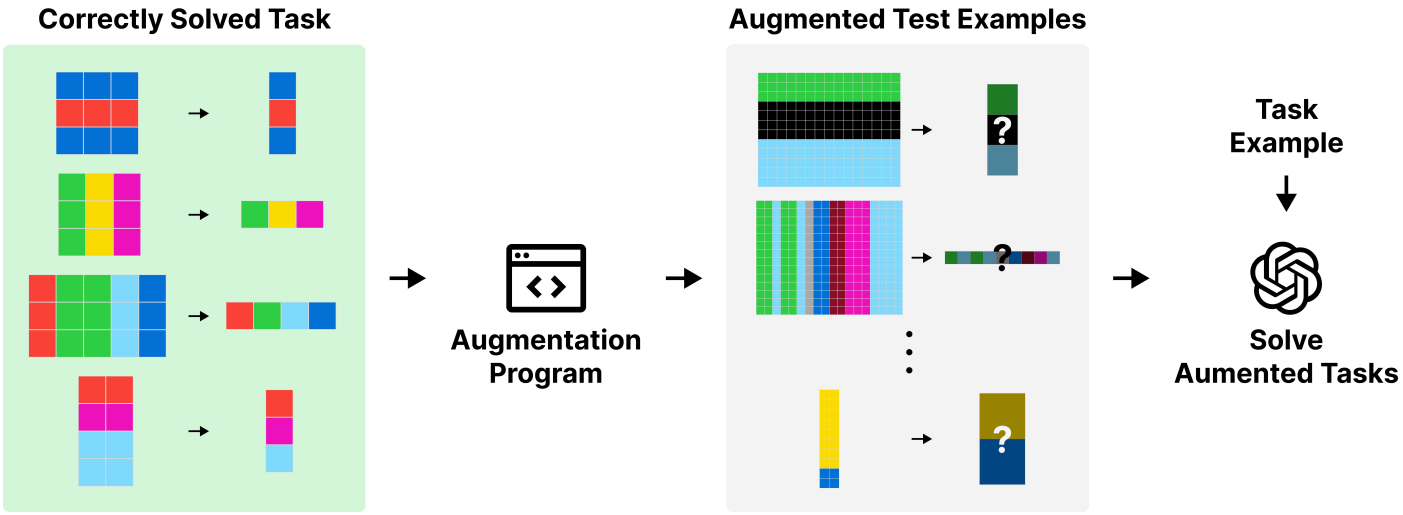
* ARC (AI2 Ressoning Challenge): This is a benchmark developed to fairly measure only the reasoning ability of artificial intelligence, and is characterized by inferring rules from input/output images using only 30X30 grids and 10 colors.



▲ Overview of the three core concepts of LoTH and the experiments that verified them through ARC. The small figures shown are a diagram (left) that shows what logical consistency, composition, and productivity are, and a diagram (right) that shows how the experiment was conducted using the ARC benchmark.

First, to measure logical consistency, an experiment was conducted to see whether LLM derived consistent correct answers when solving a problem. The research team created an 'augmented problem' that transformed the same problem and analyzed whether LLM maintained the same logic in the transformed problem. [Figure 2] Through this, we confirmed that the logical consistency of LLM differs depending on the prompting method*.

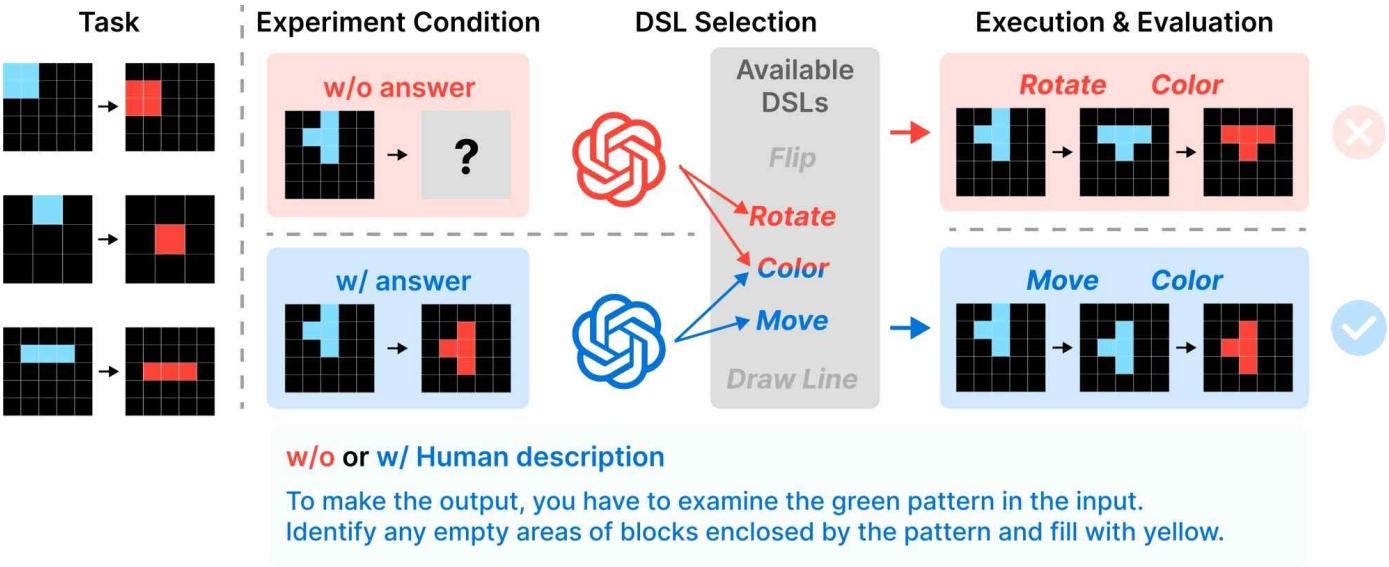
* prompting method: Instructions given to the AI model. ▲Chain of Thoughts (CoT) ▲Tree of Thoughts (ToT) ▲Least to Most (LtM), etc.



▲ Method for measuring the logical coherence of LLM and example of problem augmentation. After generating evaluation data that shares the same rules as the given task as input, LLM was measured on the augmented data.

Next, to evaluate the compositionality (combination ability), we experimented on how effectively LLM combines the concepts required to solve the problem*. Compared to humans who combine individual concepts by considering the entire process, LLM showed a decrease in accuracy as the number of steps to be combined increased.

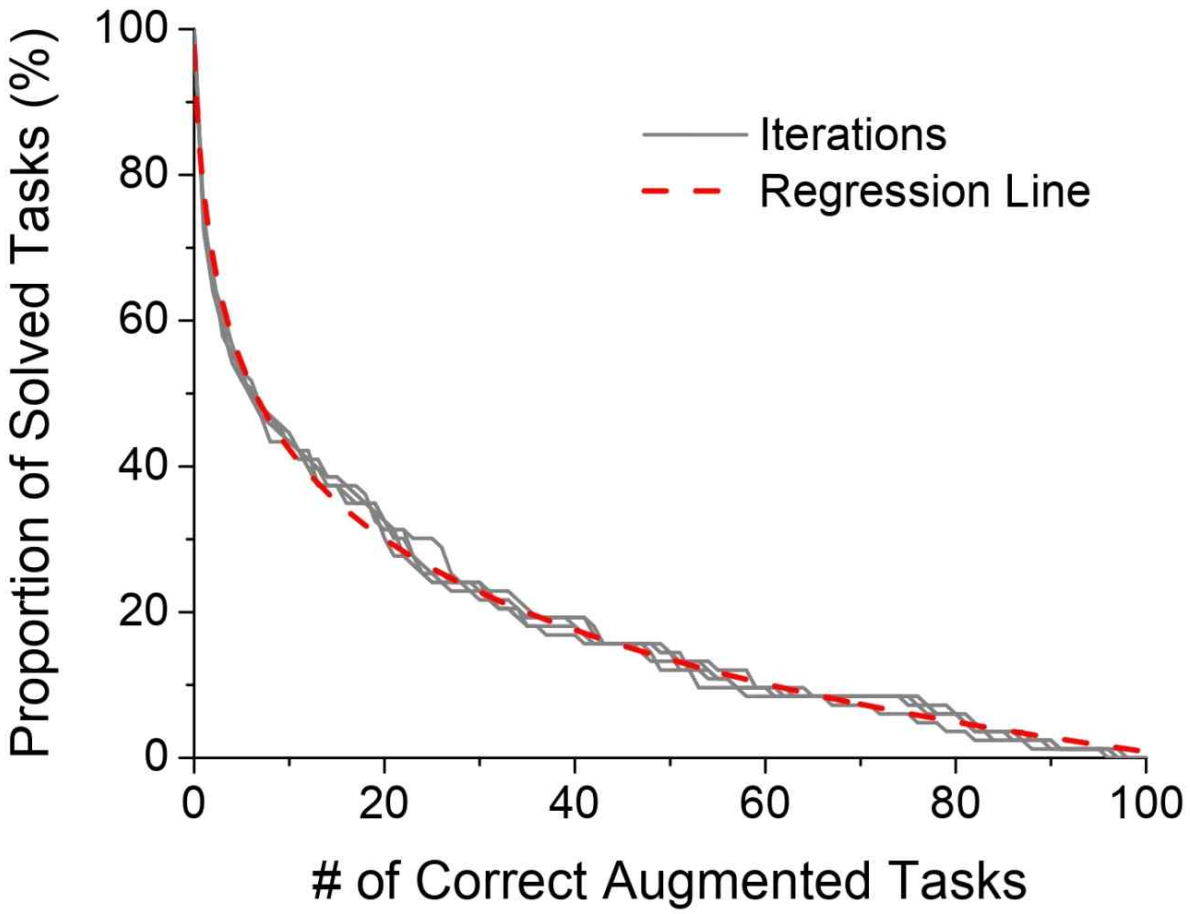
* We provided a set of subfunctions (DSL, Domain Specific Language) that can solve the ARC benchmark.



▲ Method for testing the compositionality of LLM. We first introduced a program synthesis method that selects an appropriate Python function after providing a human explanation.

Finally, to evaluate the generativity of LLM, we experimented on how many valid results that meet constraints were generated. To this end, the research team divided the ARC problem into several categories and proposed a new backward prompting method.

In addition, the research team presented an experimental method to analyze the reasoning ability of LLM from a process-oriented perspective, and in the process, proposed not only LLM but also a program synthesis method utilizing LLM necessary for the development of inference AI and a data augmentation method using a prompting technique.



▲ LLM's accuracy for augmented data. When we experimented with augmenting 100 problems for each of the 30 original tasks that LLM had solved, we confirmed that the accuracy decreased exponentially. This suggests that LLM cannot solve even if only a part of the same type of problem is transformed.

As a result of quantitatively measuring LLM's reasoning ability, it showed an average accuracy of 18.2% for augmented (transformed) problems in the logical consistency section, an accuracy of 5-15% for combination tasks in the composition section, and a generative validity of 17.12% in the generative section.

Regarding the research results, the research team explained that LLM shows some reasoning ability, but when the planning stage is long and the input/output images become complex, it cannot go through step-by-step reasoning, showing limitations in these three aspects (logical consistency, composition, and generativeness), and its reasoning ability still lags behind that of humans.

Professor Sundong Kim said, “While previous LLM evaluation methods focused on performance measurement by specific benchmarks, this study is characterized by analyzing the difference between LLM’s reasoning process and humans. We expect that it will contribute to artificial intelligence systems, including AI robots, acquiring human-level reasoning capabilities in the future.”

This study, supervised by Professor Sundong Kim of the AI Graduate School and conducted by undergraduate students Seungpil Lee, master’s students Woochang Sim, and master’s students Donghyeon Shin, was supported by the Research and Development Support Project of the Electronics and Telecommunications Research Institute, the International Joint Research Project of the National Information Society Agency for Digital Innovation Technology, and the Mid-career Researcher Support Project of the National Research Foundation of Korea, and was published online in the international academic journal 《ACM Transactions on Intelligent Systems and Technology (TIST)》 on January 20, 2025.