Reasoning Abilities of Large Language Models through the Lens of Abstraction and Reasoning

Seungpil Lee **Woochang Sim** **Donghyeon Shin** **Sejin Kim** **Sundong Kim**
Gwangju Institute of Science and Technology
{iamseungpil, dxt7469, shindong97411, sjkim7822, sdkim0211}@gist.ac.kr

Abstract

Large Language Models (LLMs) have recently demonstrated impressive capabilities across a range of natural language processing tasks. However, a fundamental question remains: to what extent do these models exhibit genuine reasoning abilities? In this study, we focus on understanding the inference processes of LLMs through an in-depth evaluation of their reasoning capabilities on tasks drawn from the Abstraction and Reasoning Corpus (ARC). Our approach takes inspiration from the "Language of Thought" Hypothesis (LoTH), which posits that human reasoning is built upon three core components: logical coherence, compositionality, and productivity. By evaluating LLMs on these three dimensions, we aim to provide insights into their reasoning strengths and limitations. Through this extended abstract, we highlight key experimental results that illuminate the capabilities and limitations of current LLMs in tasks requiring advanced cognitive reasoning. [1]

1 Motivation

Most current evaluations of LLMs focus on their output accuracy in downstream tasks such as translation, summarization, or question-answering. While these tasks are important benchmarks, they are often result-oriented and provide limited information about the underlying cognitive or logical processes of the models. Our research attempts to go deeper, examining whether LLMs exhibit coherent reasoning processes akin to human cognition. By utilizing the ARC benchmark, we push LLMs to solve tasks that require logical structuring, rule inference, and abstract pattern recognition—areas where human reasoning excels and traditional deep learning models have struggled.

The ARC dataset presents reasoning challenges that involve understanding and manipulating symbolic relationships within a grid [1]. For example, LLMs are tasked with deducing rules from a set of examples and applying those rules to a new scenario. These tasks, while seemingly simple, pose significant challenges to existing models due to their abstract nature. This abstraction demands logical consistency, compositional understanding, and productivity—precisely the dimensions we evaluate.

Typical research examines the reasoning capability of LLMs by checking how well they solve tasks like the ARC evaluation set through prompt engineering or fine-tuning [5]. However, these experiments alone make it difficult to determine whether the LLM is truly performing reasoning or simply interpolating information it has seen during training to arrive at the answer. Taking a step further, based on Fodor's Language of Thought Hypothesis (LoTH) [2], we aim to measure the reasoning capability of LLMs across three dimensions.

[1] This is an extended abstract of the paper "Reasoning Abilities of Large Language Models: In-Depth Analysis on the Abstraction and Reasoning Corpus" [4]. https://arxiv.org/pdf/2403.11793

The 1st Workshop on System-2 Reasoning at Scale, NeurIPS 2024.

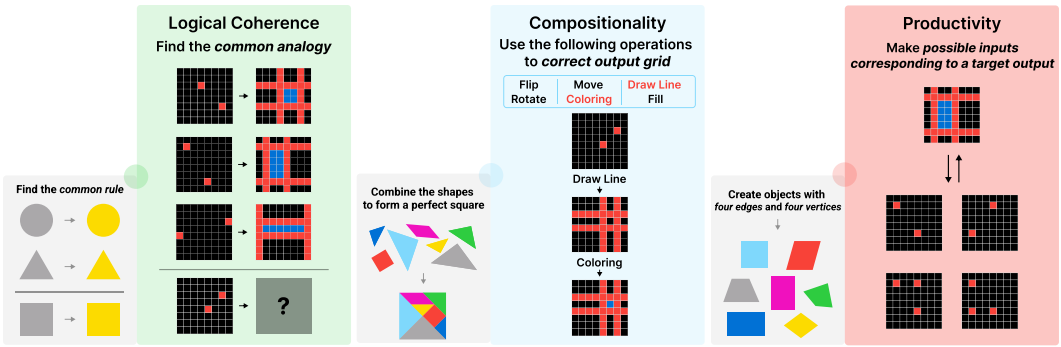


Figure 1: Three concepts of the Language of Thought Hypothesis (LoTH): logical coherence, compositionality, and productivity. These three principles serve as a foundation for evaluating the reasoning capabilities of LLMs in tasks such as ARC. By analyzing the model's ability to maintain common analogical rules, combine operations, and generate novel solutions, we can gain deeper insights into its strengths and limitations in performing human-like reasoning.

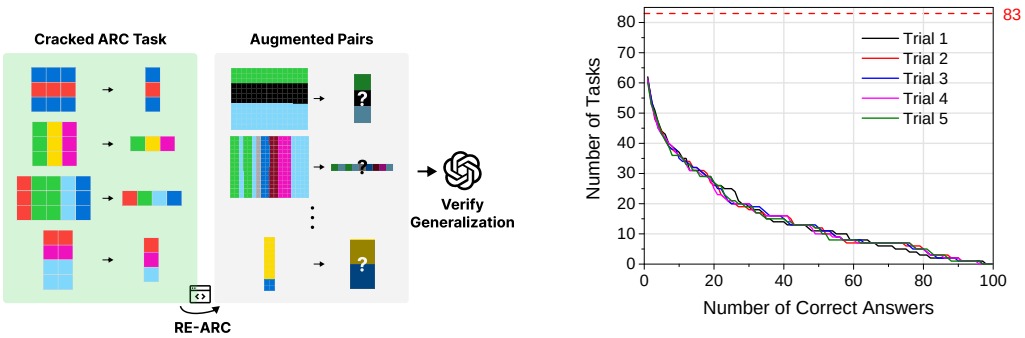
2 Evaluation of LLMs across Language of Thought Hypothesis

2.1 Logical Coherence

This component measures whether the LLM can maintain consistency in its reasoning across multiple steps. Logical coherence is essential for any reasoning process that involves drawing conclusions from premises. In the context of ARC, we evaluate whether the LLM can apply a specific logical rule consistently across multiple, related instances.

Using Re-ARC [3], we augmented 100 new test pairs for each task that GPT successfully solved. These augmented test pairs preserved the original analogical rule, allowing us to consistently assess LLM's ability to apply the rule with inferential coherence across varied instances.

Results show that while LLMs can sometimes arrive at the correct solution for an ARC task, their reasoning process is often flawed. We observed that LLMs could solve some tasks correctly but for the wrong reasons, indicating a lack of semantic coherence. Additionally, when presented with multiple test instances requiring the same rule, LLMs exhibited poor generalization performance, often failing to apply the same logical reasoning across variations of the task.



(a) Coherence measurement framework via augmentation. Using Re-ARC, we generated augmented pairs from cracked ARC tasks to verify generalization.

(b) The number of tasks completed is based on the number of correct answers out of 100 augmented test examples across five repeated trials.

Figure 2: Test performance on 100 augmented examples for each of the 83 tasks previously solved by the LLM. (a) shows the coherence measurement framework via augmentation. (b) illustrates the number of tasks completed based on correct answers across five trials.

2.2 Compositionality

Compositionality refers to the ability of a system to combine simple ideas or functions into more complex ones. For instance, in ARC tasks, LLMs must often combine simple transformations such as rotating an object or flipping it vertically to solve more intricate problems. Our evaluation focuses on whether LLMs can correctly identify and apply multiple operations in the right sequence.

To measure compositionality, we provided LLM with information about Domain-Specific Languages (DSLs) and asked them to solve given ARC tasks. Fig 3 illustrates the structure of the entire experiment. If an LLM possesses sufficient compositionality, it should be able to select appropriate DSLs and their arguments for a given goal. Four experimental conditions were tested: DSLs only, DSLs with correct output, DSLs with human descriptions, and DSLs with both correct output and human descriptions. The study used 158 ARC tasks solvable within 10 DSL steps, with each experiment repeated 10 times. GPT-4 was used as the LLM, and it was provided with ARC explanations, DSL function code, usage examples, demonstration tasks, test inputs, and object information. The LLM's task was to select appropriate DSL steps and arguments to solve the given ARC problems, with the output verified against the correct test output.

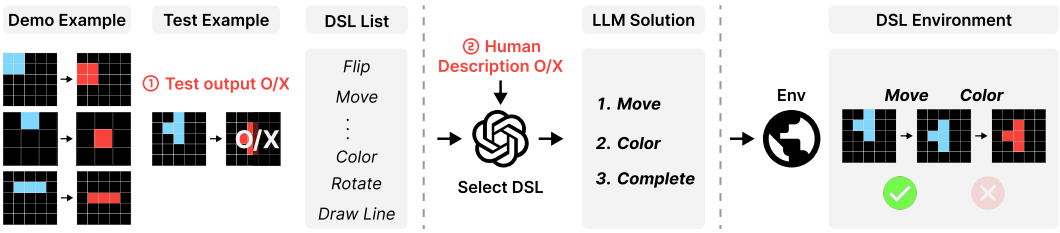


Figure 3: Overall process of compositionality experiments. Before conducting the experiment, decisions are made on whether to provide 1) the test output and 2) a human description. During execution, the LLM analyzes the given demo examples to infer the rules and then selects the appropriate DSLs from the DSL list to solve the test example. The chosen DSLs are then applied to the test input grid within the DSL environment, which determines whether the answer is correct.

The experiments showed that LLMs achieved low accuracy in solving ARC tasks using DSL, with only 3% accuracy when given just the DSL, and 9% when also provided the correct output. Including human descriptions improved performance slightly, reaching 8% without the test output and 14% with it. These results were significantly lower than human performance (86%). Further analysis revealed that LLMs could predict the output grid with 81% single-step accuracy when given the DSL and input grid, but this ability decreased as the number of steps increased. Table 1 shows the performance estimates when the single step accuracy p increases to 100%. The study concluded that LLMs struggle with both inferring rules to predict correct outputs and selecting appropriate DSLs to reach expected outputs, indicating limitations in their compositional abilities.

Table 1: The table of results shows the accuracy estimates, assuming that the LLMs have a 100% understanding of DSL, meaning the single-step accuracy p is 1.0. In reality, p was 0.81 and the results were 3%, 8%, 9%, 14% respectively (same increasing order with the table).

	w/o Human Description	**w/ Human Description**
w/o Test Output	5%	15%
w/ Test Output	17%	29%

2.3 Productivity

Productivity, in the context of the Language of Thought Hypothesis, refers to the ability to generate novel solutions or representations from a finite set of rules or elements. In human reasoning, this allows for the creation of new ideas or solutions that extend beyond the examples provided. For LLMs, productivity is assessed by their capacity to infer unseen patterns from existing ones and apply those patterns to generate valid new examples.

To evaluate productivity, the researchers designed an experiment using ARC tasks and a technique called Inverse Transformation Prompt (ITP). Given an ARC task and its abstract rule, LLMs were asked to generate valid examples of the given task. The experiment used 160 ARC tasks classified by ConceptARC [5], spanning 16 distinct categories. LLMs were provided with example pairs from the ARC task and descriptions of abstract rules applicable to similar tasks. The ITP instructed LLMs to generate multiple valid inputs that could form pairs with the output from one example of the task.

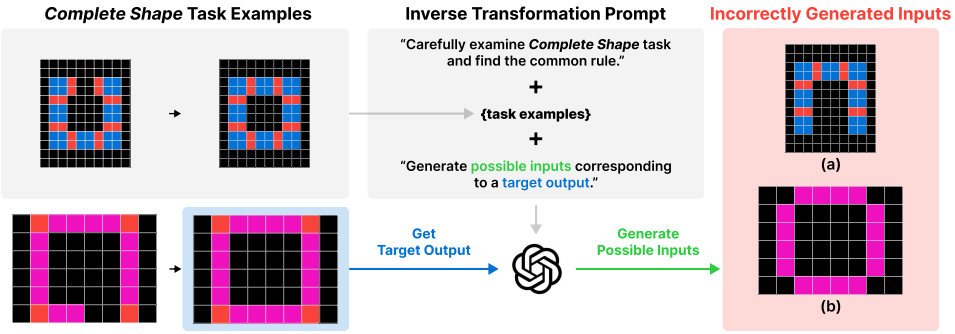


Figure 4: Two examples of the wrong generations for the task of completing the square shape. (a) LLM creates this input from the output of another example. (b) It is impossible to infer the color of the corners of the square based on this input.

The experimental results showed that LLMs struggled with productivity in the context of ARC tasks. Out of 2,913 generated examples, only about 17.1% were deemed valid according to human judgment. LLMs often failed to infer meaningful rules from given example pairs, instead resorting to simply copying inputs. They also struggled to properly consider the steps needed to generate inputs from given outputs, often creating examples that could not be solved by the specific rules of the task. These results suggest that LLMs lack a deep understanding of the semantics applicable in ARC tasks and the ability to compose these semantics according to constraints.

3 Conclusion

This paper highlights significant gaps in the reasoning abilities of Large Language Models (LLMs) as evidenced by their performance on Abstraction and Reasoning Corpus (ARC) tasks. While models like GPT-4 have shown proficiency in language tasks, they struggle with tasks requiring logical coherence, compositionality, and productivity. LLMs often arrive at correct solutions for the wrong reasons, fail to consistently apply logical rules across task variations, show low accuracy (3–14%) in combining simple operations to solve complex problems, and struggle to generate valid new examples based on given rules (only 17.1% deemed valid). These limitations are critical obstacles to scaling AI systems for more complex cognitive tasks. Our findings contribute to the ongoing discussion on how to advance scalable AI systems toward true human-like reasoning, particularly in areas requiring abstract thinking and generalization.

References

[1] François Chollet. On the Measure of Intelligence. *arXiv:1911.01547*, 2019.

[2] Jerry A Fodor. *The Language of Thought*. Harvard University Press, 1975.

[3] Michael Hodel. Addressing the Abstraction and Reasoning Corpus via Procedural Example Generation. *arXiv:2404.07353*, 2024.

[4] Seungpil Lee, Woochang Sim, Donghyeon Shin, Sanha Hwang, Wongyu Seo, Jiwon Park, Seokki Lee, Sejin Kim, and Sundong Kim. Reasoning Abilities of Large Language Models: In-Depth Analysis on the Abstraction and Reasoning Corpus. *arXiv:2403.11793*, 2024.

[5] Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The ConceptARC Benchmark: Evaluating Understanding and Generalization in the ARC Domain. *Transactions on Machine Learning Research*, 2023.