Reasoning Abilities of Large Language Models: In-Depth Analysis on the Abstraction and Reasoning Corpus

SEUNGPIL LEE*, Electrical Engineering and Computer Science, GIST, Republic of Korea
WOOCHANG SIM*, AI Graduate School, GIST, Republic of Korea
DONGHYEON SHIN*, AI Graduate School, GIST, Republic of Korea
WONGYU SEO, Electrical Engineering and Computer Science, GIST, Republic of Korea
JIWON PARK, AI Graduate School, GIST, Republic of Korea
SEOKKI LEE, AI Graduate School, GIST, Republic of Korea
SANHA HWANG, AI Graduate School, GIST, Republic of Korea
SEJIN KIM, AI Graduate School, GIST, Republic of Korea
SUNDONG KIM, AI Graduate School, GIST, Republic of Korea

The existing methods for evaluating the inference abilities of Large Language Models (LLMs) have been results-centric, making it difficult to assess the inference process. We introduce a new approach using the Abstraction and Reasoning Corpus (ARC) dataset to evaluate the inference and contextual understanding abilities of large language models in a process-centric manner. ARC demands rigorous logical structures for problem-solving, making it a benchmark that facilitates the comparison of model inference abilities with humans. Experimental results confirm that while large language models possess weak inference abilities, they still lag in terms of logical coherence, compositionality, and productivity. Our experiments highlight the reasoning capabilities of LLMs, proposing development paths for achieving human-level reasoning.

Contents

*The first three authors contributed equally and should be considered co-first authors.

1 INTRODUCTION

Recent Large Language Models (LLMs) have demonstrated performance levels close to that of humans, but experimental results showed that they lacked planning ability through thought or reasoning [7]. Consequently, a key question in recent language model research is: Can LLMs think? To address this question, new benchmarks for measuring reasoning abilities such as MathVista [39], Bongard-Logo [48], and Raven [82] have been proposed. Among them, the Abstraction and Reasoning Corpus (ARC) [9] emerged to be one of the representative benchmarks for assessing reasoning abilities. As shown in Fig. 1 below, each task in ARC consists of 2–5 demonstration example pairs and a test example input grid. The goal is to infer rules from given example pairs and apply them to the test example. Input and output grid size can vary from a minimum of 1×1 to a maximum of 30×30, with each grid having up to 10 different colors.

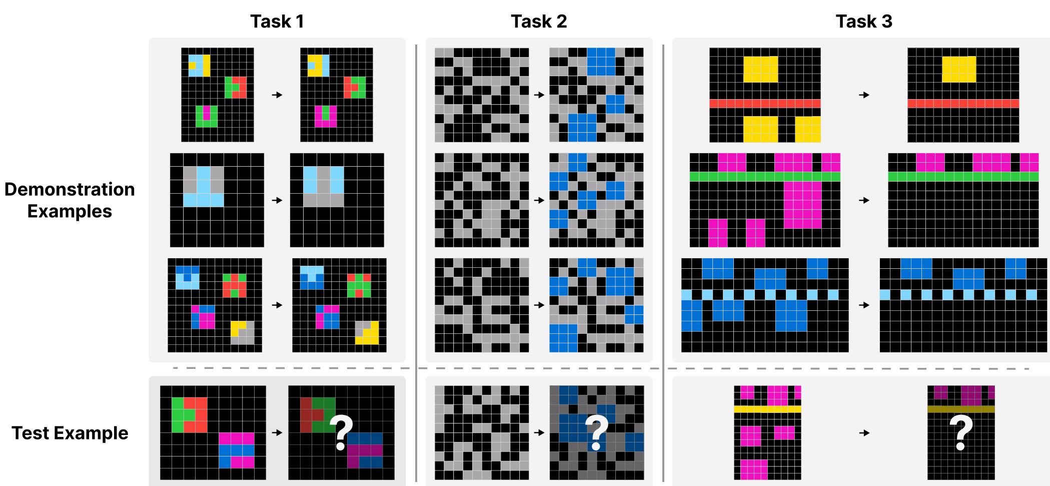


Fig. 1. Three different ARC tasks. Each task involves demonstration examples of input and output grids that exemplify the required transformation. Solvers must generate the correct output grid for the test example's input grid with the same proper transformation. ARC is a simple benchmark that can be solved using only four types of prior knowledge: objectness, goal-directedness, arithmetic, and geometric topology. Despite the small amount of prior knowledge required to solve the task, it presents a high level of reasoning difficulty. These characteristics enable ARC to function as a benchmark that fairly measures reasoning abilities.

The ARC remains an unsolved challenge despite its seemingly simple content and evaluation methods. It demands a high level of abstraction and multiple reasoning steps, reasons why conventional deep learning techniques have not achieved success. The best-performing models to date have only achieved an accuracy of 3-40% [32], while LLMs (GPT-4, PaLM) have shown an accuracy of around 10-20% [46]. Compared to the average human accuracy of 80% [29], these results suggest significant differences in reasoning and abstraction capabilities between humans and LLMs. However, in-depth research into how LLMs reason and how their reasoning differs from humans is lacking. This has led to calls for a shift from a results-focused evaluation to a more nuanced analysis of the process [2, 8, 26, 79], indicating a need for a new perspective that evaluates reasoning abilities based on the process rather than just the outcome.

To overcome the limitations of result-oriented analysis in artificial intelligence, this study adopts an existing theory on what constitutes a human's reasoning ability. According to the Language of Thought Hypothesis (LoTH) [18], human reasoning encompasses three essential characteristics:

Logical Coherence, the ability to maintain consistency in reasoning; **compositionality**, the capability to construct complex ideas from simpler components; and **productivity**, the capacity to formulate an indefinite number of thoughts or solutions using a finite set of elements.

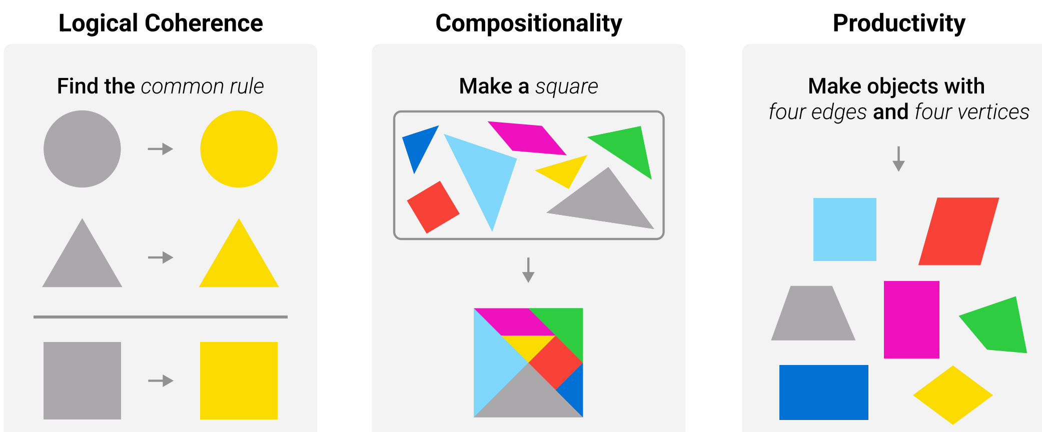


Fig. 2. Three concepts of the Language of Thought Hypothesis (LoTH).

While attempts to evaluate logical coherence, compositionality, and productivity have existed before [7, 62], there were limitations in that the definitions of each component varied across papers and existing benchmarks showed insufficient performance in assessing each aspect. This study differs from previous research in two key ways: 1) by redefining concepts borrowed from psychology to fit the field of computer science, and 2) by evaluating all elements through the visual reasoning benchmark ARC. To achieve these goals, we have designed three separate experiments:

(1) **Logical Coherence:** LoTH identifies two types of coherence. These are inferential coherence — the ability to apply logical reasoning across related instances coherently — and semantic coherence — the ability to maintain logical coherence in the reasoning process and results [19]. To verify both types of logical coherence, we augmented each solved ARC task with 100 similar test examples and evaluated the LLM's performance on these related instances. Additionally, we analyzed the solution processes, identifying cases where correct answers were derived from flawed reasoning, to measure the LLM's semantic coherence.

(2) **Compositionality:** Compositionality refers to a system's capacity to express one proposition is inherently linked to its ability to express related propositions [19]. In this study, we define compositionality as the ability to combine given semantics. Therefore, to evaluate compositionality, it is necessary to verify whether semantics can be combined as desired. Consequently, this study provided LLMs with step-by-step functions and examined whether they could identify the appropriate functions to solve ARC problems. Subsequently, we conducted an additional analysis to determine if the LLM could accurately predict the results from the given step-by-step functions and to understand the reasons for the failure.

(3) **Productivity:** Productivity refers to the ability to infinitely create unseen expressions by combining a limited set of semantics [19]. However, it is difficult to quantitatively measure whether one can make an infinite number of unseen expressions. Therefore, previous studies have evaluated productivity by assessing whether rule-compliant unseen expressions can be created [27, 33, 63]. Similarly, in this study, to evaluate the ability to generate unseen expressions, we examined whether unseen ARC tasks that comply with the rules could be generated when given a set of functions.

As a result, we have confirmed that the current level of LLM possesses a basic understanding of images and is capable of simple types of compositional object manipulations. However, compared to human reasoning abilities, LLMs lag in three areas: 1) It is not inferentially and semantically coherent. 2) Its logical reasoning abilities, especially in a step-by-step manner, are weak. 3) It struggles with understanding and generating unseen representations under complex constraints.

Finally, this study summarizes and presents recent trends proposed to address the weaknesses in abstraction abilities and reasoning capabilities. Analyzing the reasoning abilities of LLMs according to the components of human reasoning and discussing how to enhance each component represents a differentiated approach from previous research. It offers a fresh perspective for measuring and advancing the reasoning capabilities of LLMs in the future.

2 PRELIMINARIES

This section aims to explain why we chose the LoTH perspective and ARC before starting a detailed evaluation of LLM's reasoning capabilities. First, we will look at existing definitions of reasoning abilities and show why LoTH is useful in the perspective of measuring intelligence in Section 2.1. Then, in Section 2.2, we show that the ARC is an appropriate benchmark for studying LLMs from the perspective of human reasoning, as it 1) utilizes abstract semantics that can be generalized, and 2) is easy to modify.

2.1 Limitation on Assessing Reasoning Ability of LLMs

Efforts to evaluate LLMs' capabilities continue, underscoring strengths in image and text generation. Especially, analysis confirms LLMs possess elements of a World Model [23], indicating potential in inference tasks. Despite these capabilities, challenges in reasoning are noted [62], with errors such as distortion and incomplete reasoning highlighted [36]. Research suggests these reasoning abilities can improve through methodological adjustments. Furthermore, studies indicate that complex compositionality remains challenging [17].

The divergent claims about the reasoning abilities of LLMs stem from result-centric measurement methods. Turing was the first figure to shift the approach to inference towards consequential direction [60]. Subsequently, Wiener [74], McCulloch and Pitts [44], and Rosenblatt [53] shifted to studying methods for measuring performance rather than focusing on the process. Recently, Chollet attempted to quantify inference abilities from a consequential perspective [9]. However, these studies all focus on what reasoning can achieve using a result-oriented approach, without specifying the elements that constitute reasoning ability. West et al. [73] raised concerns about evaluating the reasoning ability of LLMs from a consequentialist perspective, as the generation capability of LLMs may not necessarily depend on comprehension abilities.

Therefore, a new perspective is needed to evaluate AI's inference processes; the Language of Thought Hypothesis (LoTH) enhances discussions by integrating reasoning components with quantitative metrics. LoTH posits that inference involves manipulating mental representations, a view that dominates the philosophy of mind due to its explanatory power over logical coherence, compositionality, and productivity observed in human cognition. These mental representations are believed to have a compositional syntax and combinatorial semantics. Our study compares with prior works and evaluates LLMs' inference capabilities through the LoTH, marking progress by assessing aspects like logical coherence, compositionality, and productivity.

Logical coherence, compositionality, and productivity of LLMs have been evaluated separately in previous research. The ability of LLMs to perform deductive reasoning and maintain consistency in that reasoning has been considered an important evaluation metric. In this context, logical coherence is generally defined as the ability to construct coherent logic when solving specific problems [83]. Meanwhile, whether LLMs possess compositional ability is one of the emerging questions. Given the lack of definition for compositionality, existing studies have not yet reached a unified opinion. Nevertheless, there have been attempts to define compositionality as the ability to understand and combine complex expressions, and to evaluate it by providing step-by-step prior knowledge necessary for solving problems requiring deductive reasoning [33]. Lastly, productivity is the ability to create new expressions from limited resources. With the emergence of generative AI, discussions are ongoing about what specific abilities or processes should be considered as productivity, and how to evaluate them. Some studies have judged this based on the accuracy and efficiency of the output generated using limited resources [27, 63]. However, these existing attempts have limitations in that the evaluation criteria and definitions vary across papers, and they have not compared the elements of reasoning to human reasoning processes.

Exploring deeper into the three perspectives of LoTH offers strong justification for improving reasoning capabilities. These principles help in developing the ability to process information and solve tasks similar to human reasoning. Logical coherence ensures LLMs can reason without contradictions, compositionality allows LLMs to adapt known knowledge to new scenarios, and productivity enhances LLMs' capacity to generate results based on given rules. Thus, adopting these perspectives of LoTH aids LLMs in achieving more human-like reasoning, enabling them to address complex problems with innovative and valid results.

2.2 Advantages of using ARC as Reasoning Benchmark

In exploring benchmarks suitable for evaluating inference abilities through the lens of the Language of Thought Hypothesis (LoTH), the Abstraction and Reasoning Corpus (ARC) emerges as a compelling candidate. ARC has two strengths compared to other reasoning benchmarks. First, it aligns with the LoTH perspective in that it can be solved through combinations of semantics. Second, it allows for easy task modification and generation, enabling flexible changes to objectives.

2.2.1 Core Properties of ARC. An important characteristic of the ARC benchmark is its requirement for extracting compositional semantics and combining them to resolve problems. This feature distinguishes ARC from other benchmarks and necessitates sophisticated problem-solving approaches. Two key research findings support this assertion:

(1) **Importance of Semantic Information:** Studies demonstrating enhanced performance through supplementary information underscore the significance of semantic content in ARC task resolution. In a notable experiment, the integration of graph-represented object information resulted in a substantial accuracy increase, nearly doubling the success rate [78]. This marked improvement emphasizes the critical role of semantic information in effectively addressing ARC challenges.

(2) **High Abstraction Level of ARC:** Comparative analyses reveal ARC's higher abstraction level relative to other benchmarks, as evidenced in Table 1. Chollet, the benchmark's proposer, contends that conventional feature extraction methodologies are insufficient for ARC, given its demand for complex shape interpretation and transformation process comprehension [9]. These findings highlight the necessity of employing advanced strategies capable of interpreting the abstract content integral to ARC problem-solving.

These observations confirm the importance of developing and implementing approaches that can effectively extract and utilize the complex, abstract information essential to solving ARC challenges. The unique nature of ARC necessitates a shift from traditional problem-solving paradigms toward more sophisticated, semantically-aware methodologies. Future research in this domain should focus on creating algorithms that can not only process visual information but also infer underlying rules, patterns, and transformations.

Table 1. Alignment of Abstract Visual Reasoning tasks with its taxonomy [45]. The tasks and their corresponding benchmarks are cataloged under the following four dimensions of the taxonomy: input shapes, hidden rules, target tasks, and specific challenges.

Dataset	Input Shapes		Hidden Rules		Target Tasks		Specific Challenges	
	Geometric Shapes	Abstract Shapes	Explicit Rules	Abstract Rules	Classify	Generate	Domain Transfer	Extrapolate
ARC [9]		✓		✓		✓	✓	✓
Mini-ARC [30]		✓		✓		✓	✓	✓
1D-ARC [78]		✓		✓		✓	✓	✓
MC-LARC [55]		✓		✓	✓		✓	✓
Sandia [43]	✓		✓		✓			
Synthetic [66]	✓		✓		✓			
G-set [42]	✓		✓		✓			
RAVEN [82]	✓		✓		✓		✓	
PGM [3]	✓		✓		✓		✓	✓
Hill et al. [24]	✓		✓		✓		✓	✓
G1-set [42]	✓		✓		✓		✓	
S1-set [42]	✓		✓		✓		✓	
MNS [85]	✓		✓		✓			
VAEC [71]	✓		✓		✓			✓
DOPT [71]	✓		✓		✓			✓

2.2.2 Flexibility in benchmark adaptation. Despite its simple rules, ARC remains a challenging benchmark with relatively low accuracy. LLMs achieve 15% accuracy [51], traditional program synthesis models reach 26% [75], while human average accuracy is 80% [29]. To address this challenge, various ARC variants have emerged, focusing on reducing either abstraction complexity or reasoning dimensions.

Three notable ARC variants are:

(1) **1D-ARC** [78]: Reduces dimensionality from 2D to 1D, simplifying complexity while retaining core knowledge. It effectively addresses object cohesion challenges, resulting in high LLM accuracy (about 90%).
(2) **MC-LARC** [55]: Adopts a multiple-choice format, transitioning from generative to selection tasks. GPT-4 showed strong performance (approximately 75%).
(3) **Mini-ARC** [30]: Limits grid size to 5x5, simplifying input while maintaining 2D generative characteristics. Performance remains challenging, similar to original ARC (around 15%).

These variations demonstrate ARC's transformation flexibility and support the necessity of composition in solving ARC tasks. As shown in Fig. 3, MC-LARC and 1D-ARC reduced reasoning step complexity, while Mini-ARC focused on reducing image complexity. The performance differences among these variants imply that reducing the need for complex transformation combinations can significantly improve results, highlighting the importance of combinatorial syntax in solving ARC.

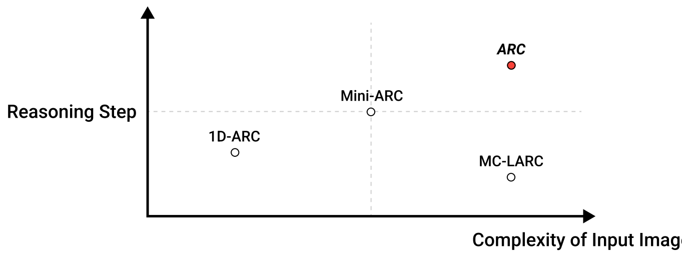


Fig. 3. Required reasoning step based on the complexity of input image in ARC, Mini-ARC, MC-LARC, and 1D-ARC. Mini-ARC and 1D-ARC simplify the task by reducing image size or dimensionality, thus lessening the reasoning needed. Meanwhile, MC-LARC changes the task format, decreasing reasoning steps but keeping the image complexity.

In summary, the ARC emerges as a compelling benchmark for evaluating inference abilities through the lens of the LoTH. ARC's core strength lies in its requirement for extracting and combining compositional semantics to solve tasks, aligning well with the LoTH perspective. This is evidenced by improved performance when additional semantic information is provided, such as object information represented as graphs. The various ARC variants (1D-ARC, MC-LARC, and Mini-ARC) demonstrate their flexibility and adaptability for different experimental purposes. The significant performance differences among these variants, particularly the high accuracy in 1D-ARC and MC-LARC compared to the original ARC and Mini-ARC, support the necessity of combinatorial syntax and sequential transformations in solving ARC tasks. Furthermore, ARC's high level of abstraction and reasoning complexity, as seen in its low accuracy rates for both LLMs and traditional models compared to human performance, underscores its value as a challenging benchmark. These characteristics collectively validate the rationale for using ARC in this study as an effective tool for exploring inference abilities in the context of the LoTH.

3 EVALUATING THE INFERENTIAL CAPABILITIES OF LLMS USING THE ARC BENCHMARK

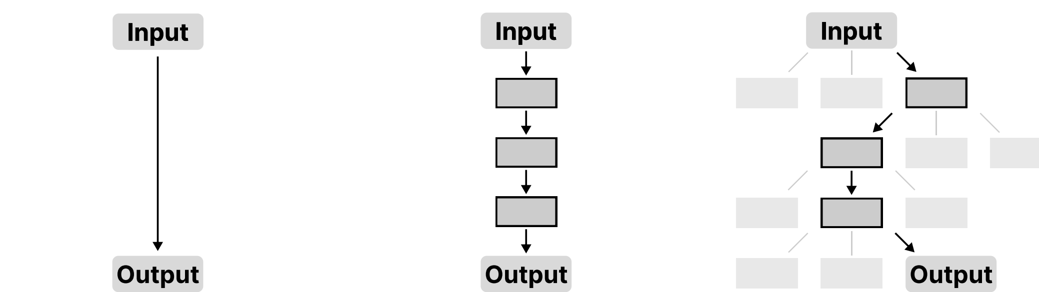
To evaluate whether LLMs possess inferential capabilities, one could compare these capabilities to human reasoning. As explained in Section 2.1, according to the Language of Thought Hypothesis (LoTH), human reasoning can be broadly divided into three main components: Logical Coherence (Section 3.1), Compositionality (Section 3.2), and Productivity (Section 3.3). We utilized Abstraction and Reasoning Corpus (ARC) to examine each aspect of the reasoning capabilities of LLMs from the perspective of LoTH.

3.1 Capability of LLMs 1: Logical Coherence

3.1.1 Motivation. Section 3.1 aims to evaluate the logical coherence of Large Language Models (LLMs). This is a fundamental aspect of the Language of Thought Hypothesis (LoTH), which considers coherence in two senses: inferential coherence and semantic coherence [19]. **Semantic coherence** refers to the ability to maintain logical consistency in the process and results of reasoning. **Inferential coherence**, on the other hand, is a system's ability to consistently apply a specific type of logical inference across all relevant instances, given it can perform that inference in some cases. These concepts are crucial in human cognitive processes and relevant to the rule inference required in ARC tasks.

Our initial experiments primarily focused on measuring semantic coherence by evaluating whether the results produced by LLMs logically followed their problem-solving steps. This was done using various prompt techniques such as Chain of Thought (CoT) [72], Least to Most (LtM) [86], and Tree of Thought (ToT) [81], similar to previous ARC solving attempts [46, 78]. We compared the coherence levels these different prompting strategies achieved, aiming to identify which techniques yielded the most semantically coherent results across diverse problem-solving scenarios. However, recognizing the limitations of this approach in addressing inferential coherence, we introduced supplementary experiments using augmented ARC tasks. These tasks, created through the Re-ARC program [25], allow us to assess how consistently LLMs can apply logical patterns across variations of originally solved problems, providing a more comprehensive evaluation of their logical reasoning capabilities.

3.1.2 Comparison Across Prompting Techniques. The perceived deficiency in LLMs' logical reasoning has been a recurrent critique, with direct attempts at solving ARC tasks yielding success rates below 10% [46]. To mitigate this, enhancements in LLMs' logical reasoning are being pursued through prompting techniques like Chain of Thought (CoT) [72], Least to Most (LtM) [86], and Tree of Thought (ToT) [81]. These strategies have been shown to leverage LLMs' reasoning capabilities [65] effectively and have the advantage of allowing for a more transparent analysis for humans, as they involve a step-by-step reasoning process. Therefore, in this experiment, we assess the impact of these prompting strategies on LLMs' logical coherence by solving ARC tasks.



(a) Chain of Thought (CoT) (b) Least to Most (LtM) (c) Tree of Thoughts

Fig. 4. Three prompting techniques in the experiment about logical coherence: (a) CoT, (b) LtM, and (c) ToT.

We applied three major prompting techniques – Chain of Thought (CoT), Least to Most (LtM), and Tree of Thought (ToT) – to solve 100 ARC evaluation tasks using the GPT-4-32k model. Each technique was tested in five iterations. ARC tasks follow a few-shot learning paradigm, requiring the model to discern task rules from given example pairs and apply them to test examples. The CoT method enhances reasoning performance by utilizing a chain of thought, providing examples of ARC task solutions in the prompt. Similar contextual information was provided for LtM and ToT. LtM decomposes tasks into manageable steps and executes them sequentially, while ToT generates multiple candidates at each step post-decomposition, selecting the best candidate through voting before proceeding to the next step.

Comparing ARC accuracy across prompts, CoT demonstrated superior performance. Table 2 presents the results of applying LtM, CoT, and ToT to 100 randomly selected tasks from the ARC evaluation set. The experiment was repeated five times, with the percentage of correct answers included for each iteration. CoT achieved approximately 10% accuracy, while LtM and ToT showed about 6% accuracy. The superior performance of CoT may be attributed to the cumulative error propagation in ToT and LtM, where small mistakes in one step of their multi-step answer generation process can lead to compounded errors in subsequent steps. Based on these results, we exclusively used the CoT prompt in subsequent experiments.

Table 2. Averaged performance of each prompting technique. The accuracy is based on solving 100 random ARC tasks with CoT, LtM, and ToT prompts, each repeated five times. The accuracy outside the parentheses refers to the accuracy when only the results are correct, while the accuracy inside the parentheses indicates the accuracy when both the results and the process are correct.

Iteration	CoT	LtM	ToT
1	11% (3%)	6% (4%)	7% (3%)
2	10% (2%)	7% (4%)	4% (1%)
3	10% (5%)	6% (3%)	7% (2%)
4	10% (4%)	4% (2%)	7% (4%)
5	10% (6%)	5% (2%)	6% (2%)
Average	10.2% (4.0%)	5.6% (3.0%)	6.2% (2.4%)

However, when we checked the correctness of the solution process, all three prompting techniques showed low accuracy, with no significant difference at around 3%, as indicated in parentheses. These results demonstrate that while there may be differences in accuracy depending on the prompting technique, there is little variation in semantic coherence. It's also important to note that both the results and processes fall far short of the average human accuracy of 80%. These findings suggest that LLMs lag behind humans in terms of logical coherence. To analyze the specific reasons for this, we conducted follow-up experiments. Section 3.1.3 analyzes inferential coherence, which is one aspect of logical coherence, while Section 3.1.4 examines the semantic coherence of LLMs through case studies.

3.1.3 Inferential Coherence of LLMs. In our second experiment, we designed a test to evaluate the inferential coherence of LLMs. Inferential coherence refers to the consistency in reasoning processes between the problem-solving steps and their outcomes. To assess this, we examined whether the LLM could solve problems similar to the ARC tasks it had previously solved successfully.

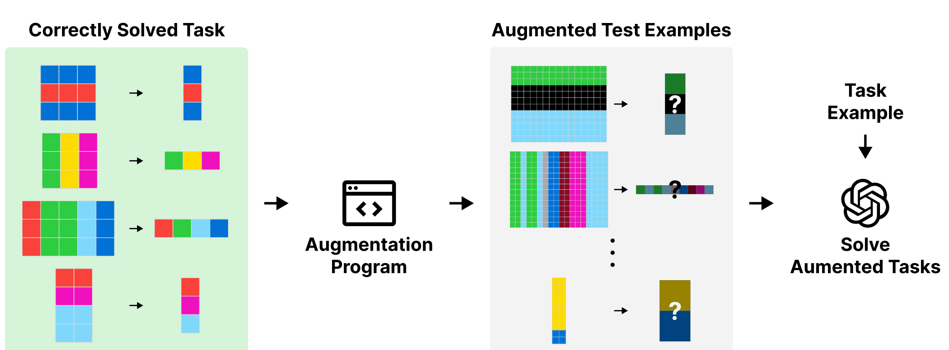
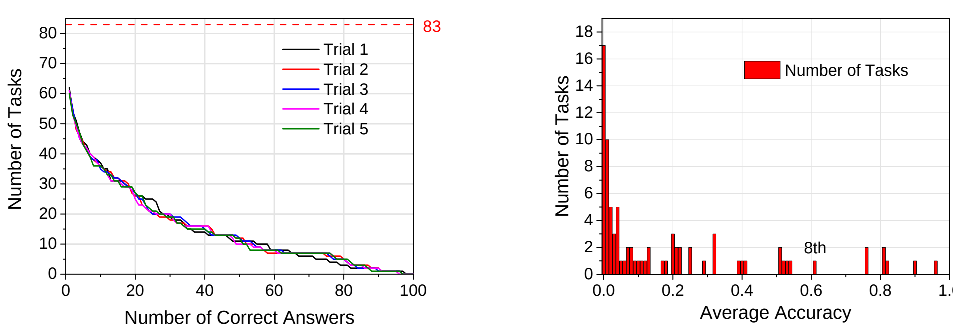


Fig. 5. Inferential coherence testing via augmentation: Using Re-ARC augmentation program [25], we generated 100 new test examples for each task that GPT successfully solved. These augmented test examples preserved the original analogical rule, allowing us to assess LLM's ability to consistently apply the rule with inferential coherence across varied instances.

Fig. 5 provides an overview of the entire experiment. We began by using GPT-4o to solve test examples from 400 ARC tasks,[1] repeating this process five times to identify which problems the model could consistently solve. For tasks that the LLM solved correctly at least once out of the five attempts, we employed Re-ARC [25] to generate 100 additional test examples for each task. These augmented test examples are methodologically identical to the original tasks regarding their solution approach. We reasoned that if the LLM truly possesses inferential coherence and applies consistent logic in problem-solving, it should be able to solve all 100 of these augmented examples. This approach allowed us to rigorously test the LLM's ability to generalize its problem-solving strategies across similar, but not identical, tasks.

Fig. 6 shows the results. Fig. 6a is a graph that represents all five iterations for the number of cumulatively correct augmented test examples. It's important to note that the graphs for all five iterations show an exponential decrease. This graph trend doesn't vary significantly between iterations, demonstrating low coherence regardless of the iteration. Fig. 6b shows the distribution of accuracy, averaged over five iterations, on 100 augmented test examples for each task. The notable points in this graph are that more than two-thirds of the tasks, 54 tasks, are concentrated below the average accuracy of 0.2, and the accuracy of the *8-th*, which corresponds to the top 10%, is about 0.6. These figures demonstrate that LLMs have low inferential coherence for most ARC tasks.



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

(b) The distribution of accuracy on 100 augmented test examples for each of the 83 tasks, averaged over five iterations.

Fig. 6. Test performance on 100 augmented examples for each of the 83 tasks previously solved by the LLM. (a) shows that LLMs failed to correctly solve 10 or more augmented test examples in over half of the tasks. (b) shows that none of the tasks correctly answered the augmented test examples and only eight tasks answered more than 60 out of the 100.

3.1.4 Case Study: Semantic Coherence of LLMs. Finally, we analyzed how LLMs solved tasks in the two experiments described in Section 3.1.2 and Section 3.1.3. When evaluating not only the answers but also the process for the three prompts CoT, LtM, and ToT, we found that regardless of the prompt, the accuracy was about 3%, indicating that correct answers were being derived from incorrect processes, as shown in Fig. 7.

To solve the task, one needs to 1) identify 5×5 objects within the input grid, 2) count the number of black squares inside each object, and 3) extract the object with the largest number of black

[1]We used 400 tasks from the ARC training set. Unlike the previous experiment, we utilized the training set instead of the evaluation set because only the training set can be augmented through Re-ARC.

squares. However, CoT, LtM, and ToT attempted to solve this task in incorrect ways. For CoT, objects appearing in the input grid were sorted first, then the object in the middle was selected. Even though CoT got the answer right, the method of sorting the objects was not comprehensible. For LtM and ToT techniques, it was understood that a specific object needed to be selected from the given input grid to solve the task. However, they mistakenly recognized objects from the test input grid. These solutions share a commonality in that they fail to explain a logically consistent rule between the provided examples of different training inputs and outputs. In other words, regardless of prompting techniques such as CoT, LtM, or ToT, LLMs have yet to demonstrate logical coherence in discovering a single rule that applies consistently across the examples provided to solve the task.

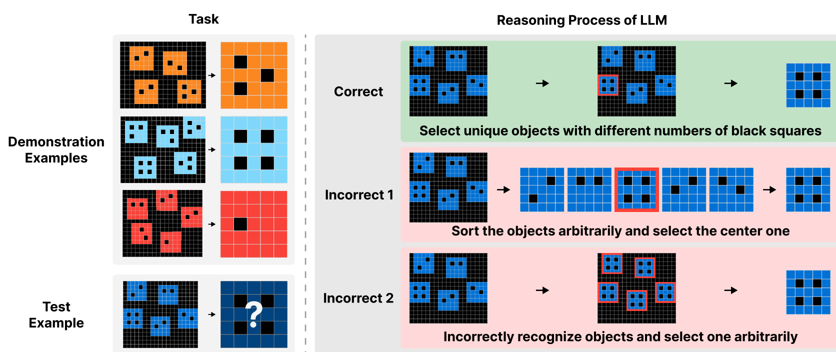


Fig. 7. Presenting instances where LLMs reach the correct answer but use flawed reasoning, highlighting the challenge of applying consistent logical rules across different ARC tasks. The task involves identifying a unique 5×5 object within a grid based on the number of black squares. The 'Correct' process shows the LLM correctly identifying the unique object, while 'Incorrect 1' and 'Incorrect 2' represent failed reasoning—one due to arbitrary selection and the other due to misidentification.

The inconsistency of inferring correct results from incorrect processes was also observed in the second experiment conducted on the training set. Upon analyzing the natural language explanations for the 83 tasks solved at least once out of 400 training tasks, we found that in 35 of these cases, the solutions proposed by the LLM could not produce the correct answer. This finding suggests that LLMs lack semantic coherence regardless of the prompting technique or the problem at hand. In other words, LLMs are deriving outcomes unrelated to their reasoning process, as evidenced by generating correct answers from incorrect solutions.

Nevertheless, in Section 3.1.3, we identified eight tasks that the LLM could solve with an accuracy of 0.6 or higher. As shown in Fig. 8, these eight tasks consists of simple solutions such as mirroring, color mapping, and partial grid copying. These tasks shared a common characteristic of conceptual simplicity, utilizing only one of the four prior knowledge domains included in ARC: objectness, goal-directedness, numbers and counting, and basic geometry [9]. For the 17 tasks that required the use of two or more prior knowledge domains, the LLM failed to solve any of the 100 augmented problems. The fact that the LLM could not solve any of the augmented problems, despite having solved the original ones, suggests that LLMs are not semantically coherent and may even indicate potential data leakage.

This comprehensive analysis demonstrates that while LLMs can solve certain simple pattern recognition tasks, they struggle with more complex reasoning that requires the integration of multiple concepts. The inability to coherently apply rules across augmented test examples, coupled with the generation of correct answers through incorrect reasoning, highlights significant limitations

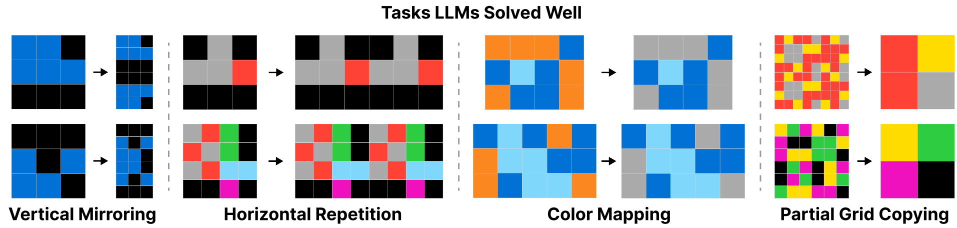


Fig. 8. Types of tasks where the LLM showed a high accuracy. The LLM showed high accuracy in simple tasks such as pattern mirroring, pattern repetition, color mapping, and partial grid copying.

in both the inferential and semantic coherence of current LLM systems when tackling abstract reasoning tasks like those presented in ARC.

3.1.5 Conclusion. In Section 3.1, we evaluated the logical coherence of LLMs by solving 100 ARC tasks using three different prompting techniques. Our results, showing accuracies ranging from 4% to 12%, demonstrate variability in reasoning performance depending on the prompting approach used. Additionally, when experimenting with GPT-4o on 400 training tasks, the LLM showed a high accuracy of 20%.

However, through an in-depth qualitative review, we demonstrated that the LLM's results may not be logically coherent. For the augmented test examples (100 for each solved task), the LLM only managed to solve 8 test examples showing performance above 60%. Furthermore, we found that for 35 out of the 83 solved tasks, nearly half, of the LLM provided incorrect solution processes that could not derive the correct results. This analysis suggests that the LLM has failed to achieve human-level logical coherence.

The results of this study align with previous research asserting that logical problem-solving remains challenging for LLMs alone. One study [64] found that LLMs can generate logically consistent reasoning with CoT prompting, even when their reasoning steps are flawed. Another study [84] showed that LLMs struggle with accurate self-reflection in tasks like mathematical reasoning and translation. Additionally, research [61] revealed that LLMs often fail to detect errors in intermediate steps, exposing flaws in their reasoning process. While these studies suggest that providing more context or enforcing stronger self-reflection might improve logical reasoning [64, 72, 84], our findings indicate that these challenges persist, suggesting the issue may not be simply a lack of information about the problem.

3.2 Capability of LLMs 2: Compositionality

3.2.1 Motivation. In Section 3.2, we investigate compositionality, the second concept of LoTH.[2] Compositionality refers to the ability to generate complex linguistic expressions given simpler ones [19]. This characteristic allows individuals to effectively tackle more complex tasks by breaking sub-tasks down into simpler steps, supporting the notion that humans can solve more complex tasks when faced with them. Strong compositionality enables the resolution of complex tasks and facilitates transparent descriptions of the process, which is also an important aspect from the perspective of LLMs.

This section uses ARC to test the compositionality of LLMs. There are previous studies that have tested a model's compositionality by providing functions in the prompt that can be combined to solve tasks, and then checking if the model can solve them [57]. Similarly, in this study, we

[2]While the Language of Thought Hypothesis principally uses the term 'systematicity', this study employs 'compositionality' as used in Fodor's paper. This choice is made because compositionality encompasses a broader concept than systemicity.

also provide step-by-step functions, which we refer to as DSL (Domain Specific Language), and then conduct experiments to verify whether they can solve ARC tasks. Additionally, to understand why tasks might not be solved, we conducted further experiments on the model's comprehension of these functions. Therefore, we verify whether LLMs understand the meaning of the functions provided for ARC tasks and whether they can combine the functions appropriately to produce the desired results. The result of this experiment indicates that while LLMs sufficiently understand the functions and their relationship with images, their ability to decompose and combine functions to achieve the desired outcome is weak.

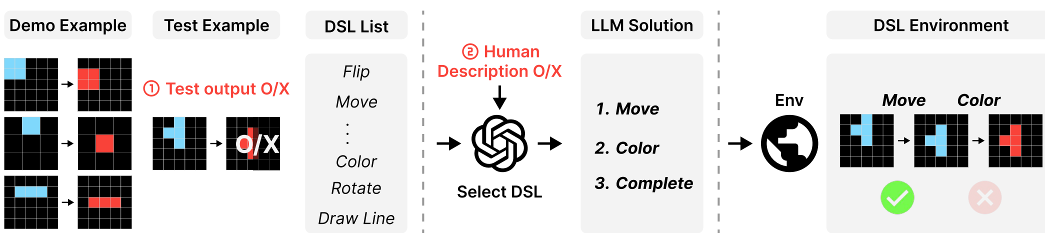


Fig. 9. Overall process of DSL compositionality experiments. Before conducting the experiment, decisions are made on whether to provide 1) the test output and 2) a human description. During execution, the LLM analyzes the given demo examples to infer the rules and then selects the appropriate DSL steps from the DSL list to solve the test example. The chosen DSL steps are then applied to the test input grid within the DSL environment, which determines whether the answer is correct.

3.2.2 Compositionality of LLMs. In the first experiment, to measure compositionality, we provided LLM with information about DSL and asked them to solve given ARC tasks. Fig. 9 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. However, in cases where the LLM failed to choose the correct DSL, we divided the conditions further to identify the cause. These conditions were whether the LLM understood the goal and whether it understood the solution process. To analyze the results according to each condition, four types of experiments were conducted: 1) given only DSL, 2) given correct output along with DSL, 3) given human descriptions to ARC test examples along with DSL, and 4) given both correct output grid and human descriptions along with DSL. Providing the correct output grid demonstrates compositionality based on knowing or not knowing the goal while providing human descriptions shows the impact of natural language descriptions on compositionality. All experiments were repeated 10 times each for 158 tasks.[3] Lastly, to establish a baseline, we also conducted human experiments. Seven participants were constrained to solve the tasks using only the same DSLs given to the LLMs.

Each DSL was given as a Python function. In this experiment, we used 19 types of DSL capable of solving ARC tasks. The prompts commonly included a brief explanation of ARC, DSL function code with comments, DSL usage examples, demonstration examples of tasks, inputs for the test examples, and object information of the test inputs. Object information is one of the crucial parameters in solving ARC tasks, which is why it was added to the prompt. We used the PnP algorithm [50] to extract object information from ARC tasks. The LLM returned a JSON-formatted string representing the chosen DSL and arguments at each step, which was used to verify whether the LLM reached the correct test output with an appropriate combination of DSL and arguments. We used the most recent model, GPT-4o, for this experiment.

[3]The 158 tasks correspond to those that can be solved within 10 steps using the given DSL, out of the total 800 publicly available ARC tasks.

Table 3. This is a table of the average accuracy from 10 repeated experiments based on the presence or absence of test output and human descriptions. The accuracy values in parentheses are Cronbach's alpha. In all the results in the table, Cronbach's alpha is greater than 0.7, indicating consistency.

	w/o Human Description	**w/ Human Description**	**Accuracy of Human**
w/o Test Output	3% (0.93)	8% (0.97)	86%
w/ Test Output	9% (0.96)	14% (0.96)	X

The experimental results are shown in Table 3. An average accuracy of 9% was observed when the test output was provided, and an average accuracy of 3% was observed without the test output. Similarly, it is an interesting finding that compositionality strengthens when human explanations are included in the prompt. As shown in Table 3, this improvement occurs at a similar rate to the test output. Cronbach's alpha measurements for each experiment showed consistency in responses, with all four experiments scoring above 0.7. However, the series of results fell significantly short of the 86% accuracy corresponding to human performance. These low accuracy rates suggest issues in the LLM's DSL composition process.

3.2.3 Analysis of compositional failures resulting from DSL misinterpretation. The issue is that the average accuracy described in Table 3 doesn't solely reflect compositionality. When we use a DSL to solve ARC tasks, we can think about the likelihood of choosing the right DSL for each step in two parts: 1) How well LLMs understand the DSL: This is reflected in how accurately it can predict the next grid when given the DSL instructions. 2) How necessary each predicted grid is in creating the final solution: This relates to how well the steps fit together to solve the task. The overall chance of picking the correct DSL for all steps depends on both of these factors working together. To solve a task, all DSLs must be correct for 10 steps. Reflecting this, we can estimate as shown in Eq. 1 below. In this equation, n represents the number of steps, w_n represents the number of tasks at step n, p represents the single-step accuracy, and x represents the difficulty of composition for each task. We assumed that the LLM's compositionality could vary depending on the information provided to the LLM and the task.

$$y = \frac{\sum_{n=1}^{10} w_n \cdot (p \cdot x)^n}{\sum_{n=1}^{10} w_n} \tag{1}$$

To determine the task accuracy considering only the compositional difficulty, we must estimate the y value when $p = 1$. Therefore, we conducted an additional experiment, as shown in Fig. 10 to verify the probability of not finding an appropriate DSL due to the inability to predict the output grid when selecting a DSL.

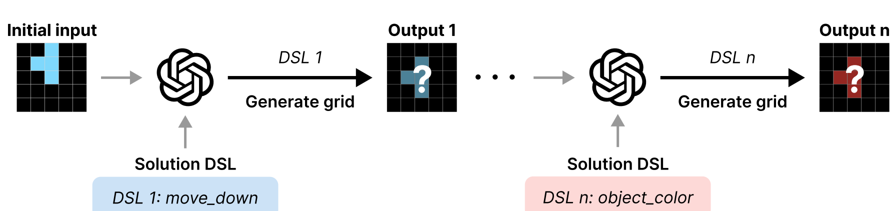


Fig. 10. Overall process of an experiment in understanding DSL. The task for the LLM is to accurately generate a grid transformed by the DSL when given a grid and its corresponding DSL. Each task involves a DSL sequence ranging from 1 to 10 steps, using trajectories previously solved by humans.

In the additional experiment, we checked how accurately the LLM could generate the correct output grid when given the DSL and ARC input grid. The experiment was conducted on 158 tasks, with each task repeated 10 times. The correct DSL and its argument chain for ARC tasks, created by humans while solving the tasks, were provided to the LLM. We prioritized using the solution with the shortest step length among the solutions provided by the human solvers. If the LLM can predict all subsequent output grids when given a specific DSL, it should be able to correctly produce the output grid regardless of the step length, since both the input grid and DSL were provided.

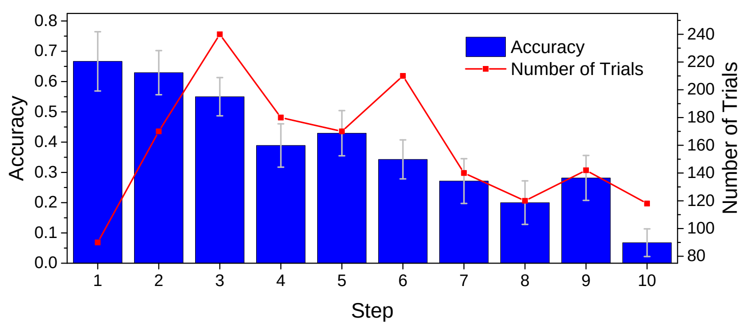


Fig. 11. Accuracy and number of tasks per step in the DSL understanding experiment. This experiment was repeated 10 times for 158 tasks. The blue bars represent the accuracy for each step, and the red line shows the number of trials for each step. As the number of steps increases, accuracy tends to decrease.

The LLM's accuracy according to the number of DSL steps needed to solve the task is shown in Fig. 11. We can observe a tendency for accuracy to decrease as the number of steps required to solve the task increases. From the above results, we used a weighted average to calculate p in Eq. 2. Eq. 2 shows the formula used to calculate p. In this equation, p represents the single-step accuracy, w_n represents the number of tasks at step n, and a_n represents the accuracy at step n. Based on this, we estimated the single-step accuracy to be 81%.

$$p = \frac{\sum_{n=1}^{10} w_n \cdot a_n}{\sum_{n=1}^{10} w_n} \tag{2}$$

Below Table 4 shows the expected modified accuracy when p corresponding to 0.8 is adjusted to 1. This table reflects only compositionality, confirming that nearly 30% of tasks can be solved when given the test output and human description. It should be noted that when both the correct answer and the natural language description leading to the answer are provided, there is a consistent increase of about 10% point. This suggests that each element positively influences the average compositional difficulty of the problem, represented as x in Eq. 1.

Table 4. The table of results shows the accuracy estimates obtained using Eq. 1, assuming that the LLMs have a 100% understanding of DSL, meaning the single-step accuracy p is 1.0.

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

3.2.4 Case Study: Enhancement of Compositionality through Human Descriptions. One notable observation was the enhanced compositionality when human descriptions of problem-solving methods were included in prompts. To investigate how LLMs could solve tasks with human descriptions, we analyzed the solution processes of 13 additional tasks solved when human descriptions were provided. Results indicate that human descriptions facilitate task input abstraction and action abstraction, thereby improving problem-solving capabilities. For instance, as illustrated in Fig. 12, LLMs fail to recognize patterns in the correct output without descriptions; however, they immediately identify patterns such as an 'X' shape with descriptions. These findings suggest the potential to enhance LLMs' reasoning performance by incorporating abstracted task information.

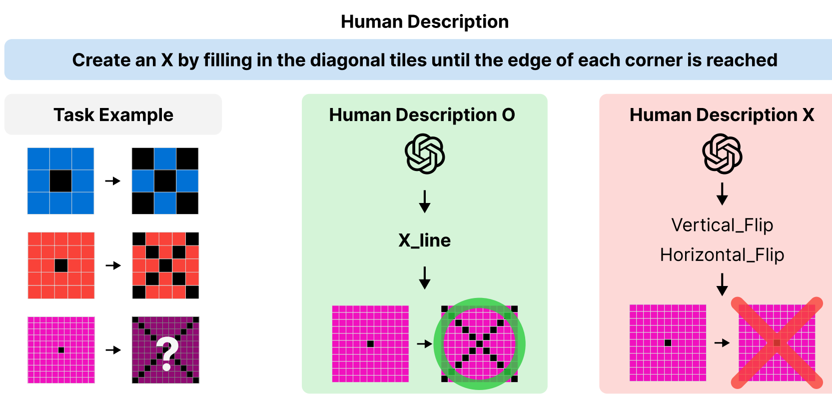


Fig. 12. Comparison of LLM's DSL selection with and without human descriptions. A human description helps find the necessary DSL for problem-solving by aiding in the required abstraction. Without it, the lack of proper abstraction prevents finding the correct DSL.

3.2.5 Conclusion. In Section 3.2, experiments using ARC and DSL were conducted to measure the compositionality of LLMs. The results led to three conclusions. First, LLMs were able to predict the output grid when DSL was applied to the input with an average accuracy of about 81%. However, as the length of steps increased, the prediction rate decreased, which appears to be due to cumulative errors. Second, when not given the correct answer, LLMs selected the correct DSL only 3% of the time, indicating a lack of ability both in inferring rules to predict the correct output grid and in selecting the appropriate DSL to reach the expected output. Finally, when human descriptions were added, the accuracy in choosing DSL increased to a level similar to when the correct answer was provided. Analysis of this process suggested that this improvement was due to linguistic abstraction of the ARC task and DSL combinations.

Previous studies have emphasized LLMs' limitations in combining simple elements to create new meanings, revealing struggles with compositionality. One study shows Transformers exhibit significant performance drops when tested on new function combinations, indicating challenges in systematically generalizing knowledge [27]. Another study introduced datasets like SADE to evaluate LLMs' ability to process visual and textual information, suggesting they still struggle with tasks like understanding negations and grasping complex content [40]. A further study examined how well LLMs can break down complex instructions or build them from simple ones. These found that while LLMs improve at understanding simple tasks by learning complex ones, they struggle with complex tasks when starting from simpler ones [80]. These findings across studies point to ongoing challenges in LLMs' ability to connect simple and complex elements, highlighting their compositionality limitations.

3.3 Capability of LLMs 3: Productivity

3.3.1 Motivation. In Section 3.3, we investigate the third concept of the Language of Thought Hypothesis (LoTH): productivity. Productivity refers to the ability to generate unseen representations based on observed data [19]. This characteristic enables humans to imagine diverse situations from a single phenomenon, facilitating efficient learning without the need for repetitive data exposure. Similarly, when endowed with this ability, LLMs are expected to excel in unseen tasks, making productivity a crucial function of essential reasoning. The capacity to generate new pairs within a constrained set of rules is particularly valuable for solving ARC tasks, highlighting the need for productivity. In this section, we will assess productivity by evaluating the validity of LLM-generated examples based on given example pairs from ARC tasks.

While productivity ideally involves testing for infinite generative capacity, practical limitations necessitate alternative approaches. The challenge lies in demonstrating that a system can produce an unlimited number of novel, meaningful outputs from a finite set of inputs and rules. Previous studies have addressed this challenge by examining whether valid outputs can be produced under added constraints [27, 33, 63]. These constraints serve to create a more manageable testing environment while still allowing for the assessment of generative capabilities. Following this methodology, our study investigates how effectively LLMs can generate valid outputs when presented with an ARC task and its underlying conceptual rule. This approach allows us to evaluate productivity within a controlled framework while still capturing the essence of generative capacity.

To understand how well LLMs can generate new expressions based on inherent logical concepts, we conduct experiments using ARC tasks. Productivity in this context involves two main steps: 1) inferring specific rules for image generation from example images and natural language expressions, and 2) applying these rules to generate new, unseen images. However, the standard approach to solving ARC tasks, as explored in previous sections, is insufficient to confirm these two processes. Therefore, we propose a novel experiment: *Given an ARC task and a basic rule shared with similar ARC tasks, can LLMs generate valid examples of the given task?* If LLMs can understand the relationship between the given ARC task and the abstract rule, they should be able to infer specific rules for the task and generate new valid examples. Through this, we aim to determine whether LLMs can imitate the productivity of human thinking in generating novel solutions.

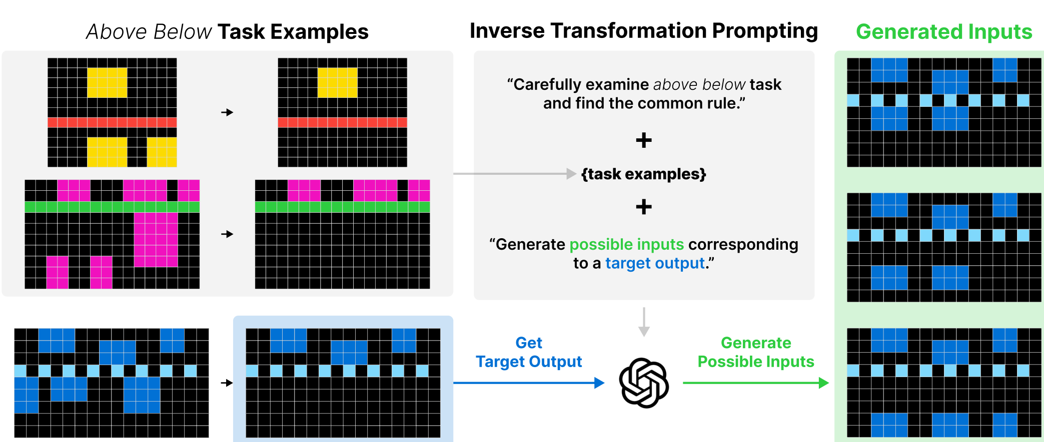
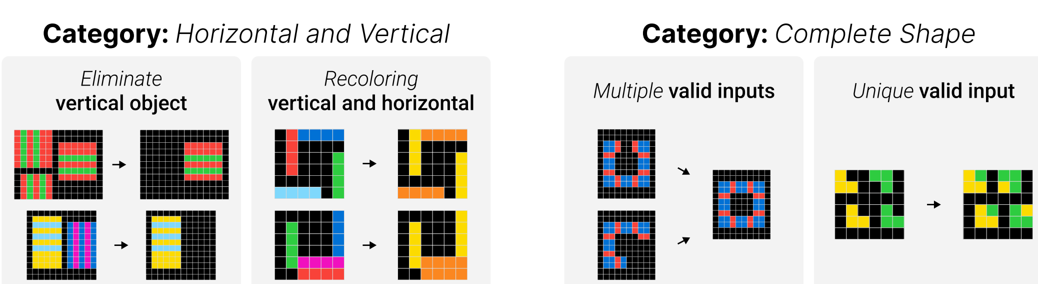


Fig. 13. Overall process of possible input generation with the Inverse Transformation Prompt (ITP). With ITP and one example of the task, LLMs generate input candidates of the output of the given example. If these generated inputs are valid, pairs created by these inputs and the given output can become new examples.

3.3.2 Validity of Augmentation. To precisely evaluate whether LLMs can infer their own generation rules given ARC examples and create new tasks by appropriately applying these rules, we rigorously controlled the prompts. LLMs receive two types of prompts: example pairs included in the ARC task and descriptions of abstract rules applicable to similar ARC tasks. However, in this case, one example pair was used as the basis for generation, and the remaining examples were used for inferring specific rules for the task. Based on the category of ConceptARC [47], which organizes a subset of ARC tasks into 16 distinct categories according to human classification criteria, we developed abstract rules. For each category within ConceptARC, we crafted a corresponding abstract rule, ensuring that tasks within the same category adhere to the identical abstract rule. An example of this abstract rule is shown at the top of the Inverse Transformation Prompt panel in Fig. 13.

We proposed the ITP, a prompting technique for this experiment. ITP instructs LLMs to generate multiple valid examples using the ARC task and its abstract rules. Fig. 13 demonstrates the process by which LLMs generate new examples given the ARC task and the corresponding ITP. LLMs generate multiple inputs that can form pairs with the output from one example of the task. This example, as mentioned earlier, is selected to serve as the basis for generation and is not included in ITP. If LLMs understood the specific rules corresponding to the ARC task given through ITP, the new example pairs generated by LLMs would be suitable as examples of the task.

ITP is based on many-to-one corresponding to elicit two advantages. First, the method of generating only input is more data-efficient than the method of generating both input and output because the output of the existing ARC task can be used as is. Since all tasks in ARC have example pairs, reusing these examples can be said to make full use of the given data. ITP allows for the reuse of a single ARC task multiple times, making it data-efficient. In particular, using ITP can further increase data efficiency by allowing one ARC task to be reused multiple times by changing the order of examples. Secondly, ITP increases the likelihood of generating valid responses. Through simple simulations, we have seen that inferring inputs from output tends to be more likely to generate valid results than inferring outputs from input. Because generating input from output is subject to relatively less stringent constraints, there is often a wide range of acceptable outcomes.



(a) Even within the same category, tasks can showcase varied objectives and complexities. The left task eliminates vertically striped objects, while the right recolors objects based on orientation.

(b) Depending on the task, there may be multiple or a unique input for an output. The left shows a task of completing a square with various inputs, and the right combines specific shapes, leading to a unique input.

Fig. 14. There are two challenges when LLMs generate examples through ITP: (a) task diversity within categories and (b) inflexibility in task-specific examples. These may cause difficulties in the process of LLMs generating examples through ITP.

In the process of creating ITP, we encounter two challenges. First, according to the ConceptARC category, there could be multiple solutions within one category. Fig. 14a illustrates that there are various types of tasks with the same category. Abstract rules given in the same sentences for each category may not be sufficient to cover various types of tasks. Second, there were ARC tasks that made not possible to infer multiple inputs from a single output (Fig. 14b). In such cases, there was only one valid input. Although we tried to take these cases into account while writing the ITP, these challenges nevertheless harmed the experimental results.

Before analyzing the experimental results, it was necessary to redefine the evaluation metric because the focus shifted from solving ARC tasks to generating valid examples. As previously explained, for a given example of a particular ARC task, we generated valid inputs that could be paired with the corresponding output. To successfully generate these inputs, the LLM must understand the specific rules of the given ARC task through its ITP and apply those understood rules to the output to create valid inputs. In this experiment, we evaluated whether all generated inputs were valid for each ARC task. This evaluation metric assesses both the LLM's understanding of the correct rules and its ability to generate valid examples based on those rules. Consequently, this experiment systematically evaluates the LLMs' capability to generate logical and valid demo pairs, enhancing our understanding of their ability to create new representations.

Table 5. The ratio of valid examples among examples generated for each category of ConceptARC.

Category	**Generated Examples**	**Valid Examples**	**Validity**
Above Below	158	34	21.52%
Center	236	35	14.83%
Clean Up	183	83	45.36%
Complete Shape	147	37	25.17%
Copy	153	4	2.61%
Count	202	29	14.36%
Extend To Boundary	167	8	4.79%
Extract Objects	176	21	11.93%
Filled Not Filled	203	29	14.29%
Horizontal Vertical	114	7	6.14%
Inside Outside	191	24	12.57%
Move To Boundary	165	12	7.27%
Order	162	26	16.05%
Same Different	246	76	30.89%
Top Bottom 2D	255	59	23.14%
Top Bottom 3D	215	25	11.63%
Total	2,913	509	17.12%

Based on 160 ARC tasks classified by ConceptARC, we evaluated the validity of a total of 2,913 generated examples. The average valid generation ratio was approximately 17.1%, with the remaining examples deemed invalid. As previously mentioned, the validity of these results was determined by human judgment, assessing whether the generated tasks adhered to the given rules. The results in Table **??** show that LLMs exhibit some capability in generating examples that align with the specified rules. However, due to weak criteria for determining validity, there is a limitation: even if an infinite number of results can be generated, they cannot be reliably used without post-processing the data.

3.3.3 Case Study: Invalid Production. An analysis of the generated inputs was conducted to investigate the reasons behind LLMs' inability to produce valid inputs for ARC tasks. Two major limitations were observed when LLMs generated new ARC tasks. First, LLMs tended to simply copy inputs rather than infer meaningful rules from given example pairs. As shown in Fig. 15, this occurred repeatedly despite attempts to prevent it through prompts. Second, LLMs failed to properly consider the steps needed to generate inputs from given outputs. This frequently resulted in the creation of examples that could not be solved by the specific rules of the task. For instance, in cases where all vertices of a square were erased in the input, it became impossible to determine the color of the vertices, making it infeasible to infer the given output. These limitations suggest that LLMs lack an understanding of the semantics applicable in ARC tasks and the ability to compose these semantics according to constraints.

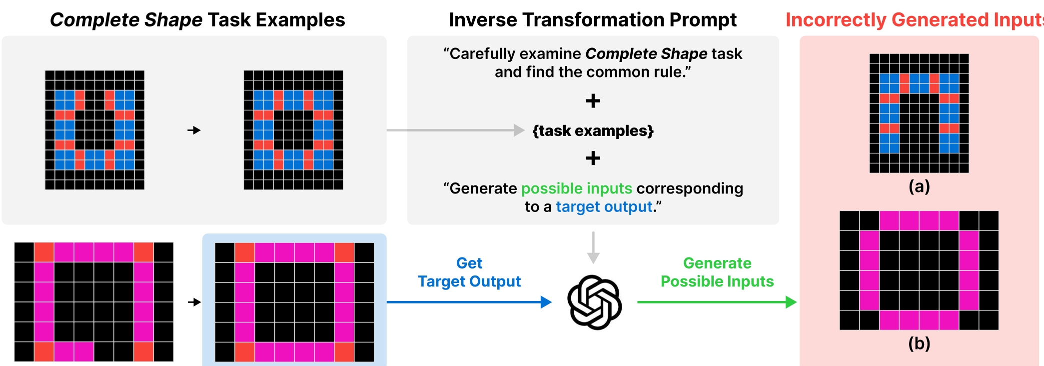


Fig. 15. Two examples of the wrong generations for the task of completing the square shape. (a) LLM creates this input from the output of another example. (b) It is impossible to infer the color of the corners of the square based on this input.

3.3.4 Conclusion. In Section 3.3, we conducted experiments to confirm the productivity of LLMs by assessing whether they can understand given ARC tasks in abstracted representations and generate valid new examples based on abstracted rules. Although it is known that LLMs have great strengths in creating creative works, our experimental results reveal that LLMs are weak in understanding rules and producing creations that adhere to those rules. Moreover, the observed limitations highlight a critical gap in LLMs' ability to engage in higher-level reasoning and abstraction, which are essential for successfully solving ARC tasks that require an understanding of underlying principles rather than surface patterns. These results suggest that when LLMs generate outputs, they tend to mimic human-created results rather than truly understanding and applying rules. This makes it difficult for LLMs to reach the level of generation that humans can achieve.

Similarly, previous studies have shown similar results in measuring the productivity of AI models. Researchers tested how well pre-LLM models generalize to novel command combinations [33, 63]. Their findings revealed strong performance on trained data but weaknesses in generating responses to unseen commands. Some researchers argued that LLMs struggle with generation under complex constraints and proposed improved models to address this issue [31, 37]. They propose novel frameworks to enhance LLMs for generating desired outputs when complex constraints are introduced, rather than relying solely on the base models. These researches share similarities with our study, which encountered difficulties in augmenting valid ARC tasks based on complex rules.

4 DISCUSSION

Through the three experiments in Section 3, we have observed that LLMs demonstrate strengths in understanding and manipulating both image and text inputs. However, they still exhibit weaknesses in logical inference, sequential planning based on understanding, and generating unseen images according to predefined rules. We will conclude by introducing the current research directions aimed at further enhancing LLMs' ability and outlining the goals after solving ARC.

4.1 What Should LLMs Possess to Solve ARC?

Based on the experimental results of Section 3, it is evident that LLMs still cannot solve ARC effectively. This is attributed to the deficiencies in logical coherence, compositionality, and productivity. How can we improve the inference capabilities of LLMs? In this section, we explore directions to enhance LLMs from the perspectives of abstraction knowledge and reasoning.

4.1.1 Abstract Knowledge. To solve ARC, the first challenge is the ability to extract the implicit information contained within ARC. Xu et al. argued that object-based representation is crucial for solving ARC and proposed ARGA [77], which converts given example grids into graphs. Their follow-up study [78] involved LLM solving ARC tasks using information obtained from ARGA and showed notable performance for object-based ARC tasks. However, these studies have a fundamental weakness in that they cannot be applied to ARC tasks without objects. Since only about 40% of ARC tasks contain object concepts [77], this method cannot be applied to more than half of the tasks. Wang et al. on the other hand, improved the abstraction ability of LLMs to some extent with a graph-form dataset consisting of 221K textual descriptions, called AbsPyramid [70], and also proposed a framework called AbsInstruct [69] utilizing this dataset. Attempting to structure sentences can be an effective abstraction method for natural language, but its effectiveness cannot be seen in tasks that do not contain sentences.

4.1.2 Reasoning. Another challenge for LLMs in the context of ARC is the vast search space. One method gaining attention to address this is to enable LLMs to generate DSLs themselves. Rajani et al. introduced CAGE [52], which prompts LLMs to generate explanations before generating answers. Subsequently, Wang et al. [67] reported improved results by having LLMs generate DSLs based on hypotheses they set themselves. Additionally, active research is underway on prompting techniques applying algorithmic approaches. Zhou et al. [87] demonstrated enhanced inference performance in LLMs by applying in-context learning. Follow-up research is actively being conducted following CoT and ToT. For example, CoT-SC [68] is a study that selects results through voting from multiple instances of CoT, GoT [4] secures flexibility by enabling the generation of graph-like thought nodes, and XoT [15] uses the thought tree while Monte Carlo tree search and refines the tree with reinforcement learning. However, these attempts are closer to additional learning for LLMs, and more researches are needed to ascertain whether fundamental improvements in LLMs' reasoning abilities are achievable.

4.2 Limitation of ARC

Does solving ARC signify the completion of human-like AI? To answer this question, two doubts need to be appropriately addressed: 1) Will the ARC solver possess human-level problem-solving abilities? and 2) Will that solver think like humans to solve ARC? It's not easy to imagine how the ARC solver operates without human-level reasoning. At this point, what we can assume is that the model will have the three properties of LoTH, and the model could be capable of several types of reasoning included in ARC. With this hypothesis, we attempt to address the following questions.

4.2.1 Will the Model Possess Human-Level Problem-Solving Abilities? Being capable of reasoning does not necessarily equate to having human-level problem-solving abilities. In other words, even if a model can reason to a level that can solve ARC, it may not have human-level problem-solving capabilities. Various tasks that humans face are generally more complex than ARC and involve various other cognitive factors besides reasoning. Therefore, even models that can solve ARC may have the following limitations compared to human-level problem-solving abilities.

First, with the current ARC criteria, it's still unknown whether the model that solved it can solve more complex types of tasks. This is because ARC tasks focus on just reasoning and are therefore presented in a relatively simple environment. Whether the reasoning ability learned through ARC would also work in more complex environments has not been revealed. Second, solving ARC does not imply the presence of other components of intelligence beyond reasoning. While reasoning is undoubtedly a core aspect of cognitive processes, it is not the entirety of intelligence. There is research shows that solving human-level complex tasks requires various cognitive abilities [21].

4.2.2 Will the Model Think Like Humans? Even if we assume that the ARC solver can reason in terms of LoTH, we cannot guarantee whether this solver's process is human-like for the following two reasons. Firstly, the current ARC provides a performance measure that rewards only for solving a task. It's important to recognize that such a measure might instigate a wrong purpose, leading to what is known as the King Midas problem [54]. This problem emphasizes the risk of AI achieving its given objective too literally, leading to unintended negative consequences, underscoring the importance of aligning AI's goals with human values and the broader context. The policy of rewarding only the results, excluding the solution process, makes it difficult to evaluate whether the solution process is similar to human reasoning. Therefore, models trained on current ARC likely differ in how they solve tasks compared to humans. The second reason is that directly comparing the reasoning processes of humans and language models is challenging. The process by which humans solve ARC tasks has not been investigated, making it unclear how the solving process differs between humans and artificial intelligence. Furthermore, there is a lack of metrics for comparing the solving processes, making direct comparisons difficult.

4.3 Future Direction After Solving ARC

To summarize, solving ARC tasks does not directly imply achieving human-level artificial intelligence. Moreover, there is a challenge in comparing task-solving approaches with those of humans. Thus, we suggest three alternatives to more accurately measure human-level inference abilities.

4.3.1 Using Different Benchmarks. One limitation of ARC is its simple environment. SQA3D [41], for instance, addresses inference tasks in a 3D domain by extending them into question-answering tasks using simulators like ScanNet [13]. Additionally, benchmarks such as TGIF-QA [28], MovieQA [59], TVQA [34], and STAR [76], which append question-answering to videos, have been proposed. Such benchmarks mimicking real-world inference scenarios could serve as supplements to measure complex abstractions not covered by ARC.

4.3.2 Quantification of ARC Task-Solving Processes. Chollet, the creator of ARC, argued that ARC maximizes generality while minimizing prior and experience [9], but these components have not been quantitatively evaluated. As a result, the quantitative assessment of factors such as the generality achieved by models solving ARC, the level of prior knowledge, and the components of prior knowledge remains elusive. One possible way to quantitatively evaluate the process of solving ARC tasks is to quantify the model's achievement of prior, experience, and generality.

4.3.3 Adding Evaluation Methods to Compare Task-Solving Processes with Human Approaches. Recent ARC research has focused on finding ways for AI to solve tasks. However, there are doubts about how similar these solutions are to those of humans. The initial paper by Johnson et al. [29] analyzed human ARC solutions. Subsequently, LARC [1] was proposed to analyze how tasks are solved through the language-based explanation of human solutions. Tools for facilitating the collection of human data are also continuously being developed. Kim et al. [30], for instance, have analyzed how tasks are solved through O2ARC. It is suggested to not only calculate simple correctness for each ARC task but also to add similarity with human data to the evaluation.

4.4 Recent Research Trends on the Reasoning Abilities of LLMs

In this paper, we utilized the ARC to evaluate and enhance the reasoning capabilities of LLMs. ARC serves as a crucial benchmark for testing AI models' ability to perform human-like reasoning, and it plays a central role in our research. Beyond ARC, various other datasets can be leveraged to strengthen LLMs' reasoning abilities further. For instance, the DROP (Discrete Reasoning Over Paragraphs) dataset [16] evaluates a model's capability to perform logical reasoning over complex text and synthesize information from multiple sources to generate accurate answers. Similarly, CommonsenseQA [58] supports training models to perform commonsense reasoning, enabling them to make logical decisions in everyday scenarios. The BoolQ dataset [11] requires models to use logical reasoning to answer true/false questions, while the GSM8K dataset [12] focuses on enhancing step-by-step logical and mathematical reasoning skills through high-quality problem-solving tasks. Together, these datasets, alongside ARC, provide invaluable resources for systematically enhancing the diverse reasoning capabilities of LLMs.

However, recent studies indicate that despite their proficiency in language-based tasks, LLMs still exhibit significant limitations in their reasoning abilities. Carvalho et al. [14] explored the reasoning capabilities of models like GPT-3.5 and GPT-4 in non-linguistic tasks requiring strategic thinking and spatial reasoning tasks. They found that while LLMs perform well on basic language tasks, they struggle with reasoning and decision-making in tasks beyond their training data, demonstrating limited cognitive flexibility and generalization. Similarly, Gendron et al. [22] critically examined the abstract reasoning skills of LLMs, revealing poor performance on tasks requiring the identification and application of general patterns from only a small number of examples, specifically designed to test broad generalization abilities. These studies collectively highlight that current LLMs, though advanced in linguistic tasks, are still far from achieving robust reasoning abilities across diverse domains.

To build on this foundation, several advanced approaches have been developed to further enhance LLMs' reasoning capabilities. Reinforcement learning with human feedback [10] allows models to learn and refine their reasoning processes through feedback during training. The Chain of Thought (CoT) prompting technique [72] facilitates multi-step reasoning by breaking down complex tasks into smaller, manageable steps. Reasoning-centric fine-tuning [35] focuses on improving performance in specific reasoning tasks by leveraging specialized datasets. Additionally, incorporating knowledge graphs during pre-training [38] enhances logical inference by integrating structured world knowledge. Finally, explainable AI techniques, such as self-reflective learning, empower LLMs to identify and correct their reasoning errors, thereby contributing to greater transparency and trustworthiness in AI [5]. Collectively, these approaches play a crucial role in advancing LLMs' reasoning capabilities across various domains.

Moreover, recent research has introduced innovative approaches to further augment the reasoning capabilities of LLMs. For example, multimodal learning techniques allow LLMs to simultaneously process text and image data, thereby enhancing their ability to tackle complex reasoning tasks [56]. Adaptive learning strategies that incorporate human feedback have also been shown to improve the models' adaptability across a wide range of tasks [49]. Additionally, integrating programming languages with LLMs has been proposed as a means to enhance logical problem-solving abilities [20]. These cutting-edge studies significantly contribute to systematically strengthening the multidimensional reasoning capabilities of LLMs.

5 CONCLUSIONS

In this study, we addressed the limitation of existing research, which predominantly analyzed LLM's inference ability from a deterministic perspective, by introducing the LoTH perspective to ensure a fair evaluation of the process as well. By comparing and analyzing arguments across various domains of inference, we confirmed that logical coherence, compositionality, and productivity are quantifiable components to evaluate inference ability. Next, we proposed three experiments using the ARC to quantitatively evaluate these three components from the LoTH perspective. The results showed that although current LLMs exhibit outstanding performance, they lack logical coherence, compositionality, and productivity in their processes, suggesting that they are closer to probabilistic mimicry rather than possessing autonomous reasoning abilities. Finally, we explored meaningful research directions for LLM to acquire reasoning abilities from the LoTH perspective, as well as alternative approaches beyond ARC. This attempt to quantitatively evaluate the inference process from the LoTH perspective through experimental methods represents a differentiated approach not present in previous research. It contributes by providing a new perspective on how to handle reasoning abilities in the field of computer science, going beyond LLMs.

ACKNOWLEDGMENTS

This work was supported by the IITP (RS-2023-00216011, RS-2024-00445087, No. 2019-0-01842) and the NRF (RS-2024-00451162) grants funded by the Ministry of Science and ICT, Korea. Experiments are supported by the Accelerate Foundation Models Research Initiative, Microsoft.

REFERENCES

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

[2] Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Physics of Language Models: Part 3.2, Knowledge Manipulation. *arXiv:2309:14402* (2023).

[3] David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. 2018. Measuring Abstract Reasoning in Neural Networks. In *ICML*.

[4] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *arXiv:2308.09687* (2023).

[5] Or Biran and Courtenay Cotton. 2017. Explanation and Justification in Machine Learning: A Survey. In *IJCAI Workshop*.

[6] Alexey Borsky. 2021. ARC-Game. https://github.com/volotat/ARC-Game

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. *arXiv:2303.12712* (2023).

[8] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology* (2024).

[9] François Chollet. 2019. On the Measure of Intelligence. *arXiv:1911.01547* (2019).

[10] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *NeurIPS*.

[11] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *NAACL*.

[12] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve

Math Word Problems. *arXiv:2110.14168* (2021).
[13] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *CVPR*. 5828–5839.
[14] Gonçalo Hora de Carvalho, Robert Pollice, and Oscar Knap. 2024. Show, Don't Tell: Evaluating Large Language Models Beyond Textual Understanding with ChildPlay. *arXiv:2407.11068* (2024).
[15] Ruomeng Ding, Chaoyun Zhang, Lu Wang, Yong Xu, Minghua Ma, Wei Zhang, Si Qin, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. 2023. Everything of Thoughts: Defying the Law of Penrose Triangle for Thought Generation. *arXiv:2311.04254* (2023).
[16] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *NAACL*.
[17] 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. 2023. Faith and Fate: Limits of Transformers on Compositionality. In *NeurIPS*.
[18] Jerry A Fodor. 1975. *The Language of Thought*. Vol. 5. Harvard University Press.
[19] Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition* 28, 1-2 (1988), 3–71.
[20] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-Aided Language Models. In *ICML*. 10764–10799.
[21] Howard E Gardner. 2011. *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books.
[22] Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2024. Large Language Models are Not Strong Abstract Reasoners. *IJCAI* (2024).
[23] Wes Gurnee and Max Tegmark. 2023. Language Models Represent Space and Time. *arXiv:2310.02207* (2023).
[24] Felix Hill, Adam Santoro, David GT Barrett, Ari S Morcos, and Timothy Lillicrap. 2019. Learning to Make Analogies by Contrasting Abstract Relational Structure. In *ICLR*.
[25] Michael Hodel. 2024. Addressing the Abstraction and Reasoning Corpus via Procedural Example Generation. *arXiv:2404.07353* (2024). https://github.com/michaelhodel/re-arc?tab=readme-ov-file
[26] Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards Reasoning in Large Language Models: A Survey. In *ACL Findings*.
[27] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality Decomposed: How Do Neural Networks Generalise? *Journal of Artificial Intelligence Research* 67 (2020), 757–795.
[28] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *CVPR*. 2758–2766.
[29] Aysja Johnson, Wai Keen Vong, Brenden M. Lake, and Todd M. Gureckis. 2021. Fast and Flexible: Human Program Induction in Abstract Reasoning Tasks. In *CogSci*.
[30] Subin Kim, Prin Phunyaphibarn, Donghyun Ahn, and Sundong Kim. 2022. Playgrounds for Abstraction and Reasoning. In *NeurIPS Workshop on nCSI*.
[31] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating Images with Multimodal Language Models. In *NeurIPS*, Vol. 36.
[32] Lab42. 2024. ARC-PRIZE competition. https://arcprize.org/
[33] Brenden Lake and Marco Baroni. 2018. Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In *ICML*. PMLR, 2873–2882.
[34] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*.
[35] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving Quantitative Reasoning Problems with Language Models. In *NeurIPS*.
[36] Zhaoyi Li, Gangwei Jiang, Hong Xie, Linqi Song, Defu Lian, and Ying Wei. 2024. Understanding and Patching Compositional Reasoning in LLMs. *arXiv:2402.14328* (2024).
[37] Mushui Liu, Yuhang Ma, Xinfeng Zhang, Yang Zhen, Zeng Zhao, Zhipeng Hu, Bai Liu, and Changjie Fan. 2024. LLM4GEN: Leveraging Semantic Representation of LLMs for Text-to-Image Generation. *arXiv:2407.00737* (2024).
[38] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: Enabling Language Representation with Knowledge Graphs. In *AAAI*.
[39] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating Math Reasoning in Visual Contexts with GPT-4V, Bard, and Other Large Multimodal Models. In *ICLR*.
[40] Teli Ma, Rong Li, and Junwei Liang. 2024. An Examination of the Compositionality of Large Generative Vision-Language Models. In *NAACL*.

[41] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2023. SQA3D: Situated Question Answering in 3D Scenes. In *ICLR*.

[42] Jacek Mańdziuk and Adam Żychowski. 2019. DeepIQ: A Human-Inspired AI System for Solving IQ Test Problems. In *IJCNN*.

[43] Laura E Matzen, Zachary O Benz, Kevin R Dixon, Jamie Posey, James K Kroger, and Ann E Speed. 2010. Recreating Raven's: Software for Systematically Generating Large Numbers of Raven-Like Matrix Problems With Normed Properties. *Behavior Research Methods* 42, 2 (2010), 525–541.

[44] Warren S McCulloch and Walter Pitts. 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics* 5 (1943), 115–133.

[45] Małkiński Mikołaj and Mańdziuk Jacek. 2023. A Review of Emerging Research Directions in Abstract Visual Reasoning. *Information Fusion* 91 (2023), 713–736.

[46] Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large Language Models as General Pattern Machines. *arXiv:2307.04721* (2023).

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

[48] Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. 2020. Bongard-Logo: A New Benchmark for Human-Level Concept Learning and Reasoning. In *NeurIPS*.

[49] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simense, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *NeurIPS*.

[50] Jaehyun Park, Jaegyun Im, Sanha Hwang, Mintaek Lim, Sabina Ualibekova, Sejin Kim, and Sundong Kim. 2023. Unraveling the ARC Puzzle: Mimicking Human Solutions with Object-Centric Decision Transformer. *ICML Workshop on ILHF* (2023).

[51] Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and Xiang Ren. 2024. Phenomenal Yet Puzzling: Testing Inductive Reasoning Capabilities of Language Models with Hypothesis Refinement. In *ICLR*.

[52] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *ACL*.

[53] Frank Rosenblatt. 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review* 65, 6 (1958), 386–408.

[54] Stuart J Russell and Peter Norvig. 1995. *Artificial Intelligence: A Modern Approach.* Pearson.

[55] Donghyeon Shin, Sanha Hwang, Seokki Lee, Yunho Kim, and Sundong Kim. 2023. MC-LARC Benchmark to Measure LLM Reasoning Capability. In *Korea Software Congress*.

[56] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A Foundational Language and Vision Alignment Model. In *CVPR*.

[57] Sania Sinha, Tanawan Premsri, and Parisa Kordjamshidi. 2024. A Survey on Compositional Learning of AI Models: Theoretical and Experimental Practices. *arXiv:2406.08787* (2024).

[58] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *NAACL*.

[59] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *CVPR*. 4631–4640.

[60] Alan Turing. 1950. Computing Machinery and Intelligence. *Mind* 59, 236 (1950), 433–460.

[61] Gladys Tyen, Hassan Mansoor, Victor Cărbune, Yuanzhu Peter Chen, and Tony Mak. 2024. LLMs cannot find reasoning errors, but can correct them given the error location. In *ACL Findings*.

[62] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2024. PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change. In *NeurIPS*.

[63] Frank van der Velde, Gwendid T van der Voort van der Kleij, and Marc de Kamps. 2004. Lack of Combinatorial Productivity in Language Processing with Simple Recurrent Networks. *Connection Science* 16, 1 (2004), 21–46.

[64] Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. In *ACL*.

[65] Haoyu Wang, Mo Yu, Xiaoxiao Guo, Rajarshi Das, Wenhan Xiong, and Tian Gao. 2019. Do Multi-Hop Readers Dream of Reasoning Chains?. In *EMNLP Workshop on MRQA*.

[66] Ke Wang and Zhendong Su. 2015. Automatic Generation of Raven's Progressive Matrices. In *IJCAI*.

[67] Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. 2024. Hypothesis Search: Inductive Reasoning with Language Models. In *ICLR*.

[68] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *ICLR*.

[69] Zhaowei Wang, Wei Fan, Qing Zong, Hongming Zhang, Sehyun Choi, Tianqing Fang, Xin Liu, Yangqiu Song, Ginny Y Wong, and Simon See. 2024. AbsInstruct: Eliciting Abstraction Ability from LLMs through Explanation Tuning with Plausibility Estimation. *arXiv:2402.10646* (2024).

[70] Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2023. AbsPyramid: Benchmarking the Abstraction Ability of Language Models with a Unified Entailment Graph. *arXiv:2311.09174* (2023).

[71] Taylor Webb, Zachary Dulberg, Steven Frankland, Alexander Petrov, Randall O'Reilly, and Jonathan Cohen. 2020. Learning Representations That Support Extrapolation. In *ICML*.

[72] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.

[73] Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2024. The Generative AI Paradox: "What It Can Create, It May Not Understand". In *ICLR*.

[74] Norbert Wiener. 1950. Cybernetics. *Bulletin of the American Academy of Arts and Sciences* 3, 7 (1950), 2–4.

[75] Johan Sokrates Wind. 2020. ARC-Solution. https://github.com/top-quarks/ARC-solution

[76] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2021. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *NeurIPS*.

[77] Yudong Xu, Elias B. Khalil, and Scott Sanner. 2023. Graphs, Constraints, and Search for the Abstraction and Reasoning Corpus. In *AAAI*.

[78] Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias B Khalil. 2023. LLMs and the Abstraction and Reasoning Corpus: Successes, Failures, and the Importance of Object-based Representations. *Transactions on Machine Learning Research* (2023).

[79] Cheng Xue, Vimukthini Pinto, Chathura Gamage, Ekaterina Nikonova, Peng Zhang, and Jochen Renz. 2023. Phy-Q as a Measure for Physical Reasoning Intelligence. *Nature Machine Intelligence* 5, 1 (2023), 83–93.

[80] Haoran Yang, Hongyuan Lu, Wai Lam, and Deng Cai. 2024. Exploring Compositional Generalization of Large Language Models. In *NAACL Workshop*.

[81] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *NeurIPS*.

[82] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. RAVEN: A Dataset for Relational and Analogical Visual Reasoning. In *CVPR*.

[83] Jinghan Zhang, Xiting Wang, Weijieying Ren, Lu Jiang, Dongjie Wang, and Kunpeng Liu. 2024. RATT: A Thought Structure for Coherent and Correct LLM Reasoning. *arXiv:2406.02746* (2024).

[84] Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024. Self-Contrast: Better Reflection Through Inconsistent Solving Perspectives. In *ACL*.

[85] Wenhe Zhang, Chi Zhang, Yixin Zhu, and Song-Chun Zhu. 2020. Machine Number Sense: A Dataset of Visual Arithmetic Problems for Abstract and Relational Reasoning. In *AAAI*.

[86] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *ICLR*.

[87] Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching Algorithmic Reasoning via In-Context Learning. *arXiv:2211.09066* (2022).

A SUPPLEMENTARY ANALYSIS

A.1 Comparing LLM and Human Problem Difficulty Perception

Following the analysis in Section 3.1.4, we analyzed problems that LLMs (Large Language Models) solve well and those they struggle with. Table 6 presents the accuracy of LLMs across problem difficulty levels classified by humans. The classification was based on the existing categorization relying on perceived difficulty by humans [6]. As a result, we discovered a tendency where problems perceived as difficult by humans align closely with those challenging for LLMs. Difficult problems shared two commonalities: 1) they required lengthy inference processes to solve, and 2) they involved considering multiple simultaneous problems to extract information about changes. An example from Fig. 16 illustrates this point: a task classified as 'Entry' only requires a single step of coloring, while a task classified as 'Hard' requires three steps: recognizing each object, identifying the priority of each object, and merging each object considering their priority. 'Easy' and 'Medium' are tasks that require relatively more complex steps than 'Entry' and fewer steps than 'Hard'. Considering these observations, it can be inferred that artificial intelligence possesses simple forms of visual logic that deal with only one of the four priors included in ARC: objectness, goal-directedness, numbers and counting, and basic geometry. However, it cannot handle complex combinations of logic that integrate these priors.

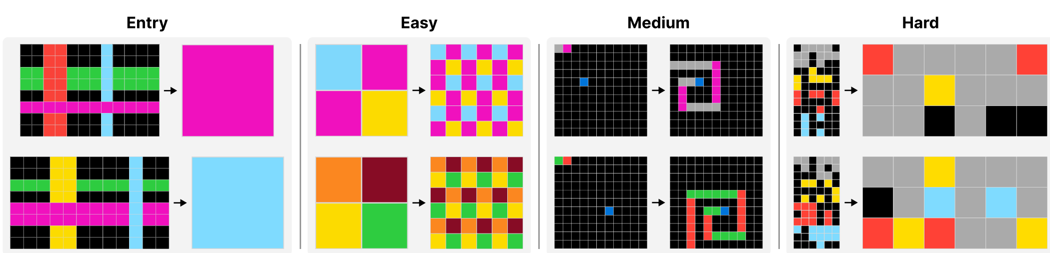


Fig. 16. Showcases of ARC tasks organized by human-perceived difficulty levels. These tasks illustrate the spectrum of complexity that humans use to rate problems, ranging from single-step 'Entry' level tasks to multi-step 'Hard' challenges. The difficulty classification reflects both the depth of inference required and the number of logical operations needed to reach a solution, paralleling the varying success rates of LLMs in tackling these tasks.

Table 6. Analyzing LLMs' reasoning capabilities by task difficulty, following prior categorization [6]. The number of ARC tasks corresponding to each category is listed in the table, and the experiment was performed five times for each task.

	Entry	Easy	Medium	Hard
Tasks	2	20	46	14
Trials	10	100	230	70
CoT	100.00%	30.00%	0.00%	0.00%
LtM	20.00%	19.00%	0.00%	2.85%
ToT	50.00%	22.00%	0.00%	0.00%
Average	56.67%	23.67%	0.00%	0.95%

A.2 Comparison of Augmentation Cost-Efficiency across GPT Versions

In a follow-up experiment to our productivity study, we aimed to compare the cost-efficiency of GPT-3.5 and GPT4-32k when augmenting demonstration example tasks. This investigation was crucial to understanding the trade-offs between model performance and associated costs in real-world applications.

Our experimental setup began with the creation of a prompt describing the category, as detailed in Table 15. Using this prompt, we developed an Inverse Transformation Prompt (ITP) and proceeded to augment demonstration examples using both GPT-3.5-16k and GPT-4-32k models. Throughout this process, we meticulously logged all prompts given to the LLMs and their corresponding responses.

To analyze the cost implications, we tokenized the logged text using the tiktoken library. We then calculated the cost of generating a single valid demonstration example based on the per-token cost specified by the Azure OpenAI API. This approach allowed us to accurately assess the financial implications of using each model for demonstration example augmentation. Validation of the generated examples was a critical component of our experiment. We employed human reviewers to manually verify the quality and appropriateness of the outputs. These reviewers were tasked with confirming two key aspects:

(1) Whether the results could be legitimately generated from the given rules.
(2) If the generated results were unique, avoid repetition or trivial variations.

This rigorous validation process ensured that our assessment of "valid" examples was thorough and meaningful in the context of practical applications.

Table 7. Comparison of augmentation cost-efficiency across GPT versions. The experiment was conducted on each demonstration example pair within the 16 task categories of Concept ARC. This table shows the results of LLMs generating valid demonstration example pairs and costs.

	Generated Example	**Valid Example**	**Validity**	**Cost per Valid Example**
GPT-3.5-16k	346	24	6.94%	\$0.0275
GPT-4-32k	411	40	9.73%	\$0.3925

Analysis of the cost to generate valid demonstration examples, as illustrated in Table 7, reveals that while GPT-4-32k showed approximately 1.5 times higher performance in terms of validity compared to GPT-3.5-16k, its cost was nearly 20 times higher. This suggests that productivity gains may not scale linearly with model capability and cost, especially when generating outputs under complex constraints. Consequently, in scenarios requiring valid outputs under intricate constraints, GPT-3.5 might be preferable to GPT4-32k when considering the trade-off between performance improvement and cost increase. However, the low overall validity rate of less than 10% for both models indicates that current LLMs still have significantly lower productivity compared to humans in such tasks. This finding suggests that merely upgrading to more advanced models is unlikely to fully resolve the productivity gap, highlighting the need for further research and development in enhancing LLM performance for complex, constrained tasks.

B EXPERIMENTAL DETAIL

B.1 Logical Coherence

The logical coherence study comprised two main experiments: a comparison on semantic coherence across prompting techniques and an assessment of the inferential coherence of LLMs. For the first experiment, prompting techniques comparison, we randomly selected 100 tasks from the ARC evaluation set. We then applied three different prompting methods - Chain of Thought (CoT), Least to Most (LtM), and Tree of Thoughts (ToT) - to compare their effectiveness in maintaining semantic coherence.

The second experiment assessing the inferential coherence of the LLMs aims to assess whether the same logic can be consistently applied. Therefore, it is necessary to first confirm the tasks where the LLMs have understood the logic. To this end, we experimented using CoT prompting, which showed the best performance in the comparison across prompting techniques experiment, to solve the ARC training set. This experiment was repeated five times. The Inferential Coherence of the LLMs experiment was then conducted on tasks that were correctly solved at least once out of the five repetitions. The detailed task IDs and prompts used in each experiment are provided in B.1.1 and B.1.2, respectively.

B.1.1 Task ID List for Each Experiment.

Table 8. Task ID list selected for the experiment comparing logical coherence. The first experiment for comparison across prompting techniques was conducted on 100 ARC evaluation tasks, while the second experiment for the inferential coherence experiment of LLMs was carried out on 83 ARC training tasks.

Experiment	**Task ID List**
Comparison Across Prompting Techniques	‘3ed85e70’, ‘0f63c0b9’, ‘17cae0c1’, ‘47996f11’, ‘4acc7107’, ‘0692e18c’, ‘477d2879’, ‘1c0d0a4b’, ‘292dd178’, ‘1990f7a8’, ‘22a4bbc2’, ‘4364c1c4’, ‘2f0c5170’, ‘17b80ad2’, ‘03560426’, ‘0c786b71’, ‘3391f8c0’, ‘42a15761’, ‘0bb8deee’, ‘1e97544e’, ‘1c02dbbe’, ‘4b6b68e5’, ‘2a5f8217’, ‘3194b014’, ‘1acc24af’, ‘0c9aba6e’, ‘0e671a1a’, ‘37d3e8b2’, ‘0becf7df’, ‘0607ce86’, ‘3a301edc’, ‘2546ccf6’, ‘009d5c81’, ‘31adaf00’, ‘281123b4’, ‘3d31c5b3’, ‘423a55dc’, ‘1d0a4b61’, ‘1a2e2828’, ‘319f2597’, ‘3979b1a8’, ‘12422b43’, ‘140c817e’, ‘0a2355a6’, ‘19bb5feb’, ‘332efdb3’, ‘27a77e38’, ‘2c0b0aff’, ‘00dbd492’, ‘2c737e3’, ‘2072aba6’, ‘48f8583b’, ‘27f8ce4f’, ‘14754a24’, ‘32e9702f’, ‘195ba7dc’, ‘137f0df0’, ‘184a9768’, ‘29700607’, ‘1c56ad9f’, ‘15663ba9’, ‘4c177718’, ‘136b0064’, ‘0a1d4ef5’, ‘1d398264’, ‘09c534e7’, ‘2685904e’, ‘48131b3c’, ‘31d5ba1a’, ‘2697da3f’, ‘103eff5b’, ‘12997ef3’, ‘1e81d6f9’, ‘25094a63’, ‘08573cc6’, ‘20981f0e’, ‘4852f2fa’, ‘2b01abd0’, ‘2072aba6’, ‘1a6449f1’, ‘34b99a2b’, ‘0b17323b’, ‘15696249’, ‘414297c0’, ‘2753e76c’, ‘12eac192’, ‘0934a4d8’, ‘310f3251’, ‘358ba94e’, ‘21f83797’, ‘4aab4007’, ‘351d6448’, ‘45bbe264’, ‘456873bc’, ‘15113be4’, ‘3490cc26’, ‘3b4c2228’, ‘00576224’, ‘42918530’, ‘45737921’, ‘20818e16’

Experiment	**Task ID List**
Inferential Coherence of LLMs	‘017c7c7b’, ‘025d127b’, ‘08ed6ac7’, ‘0dfd9992’, ‘1bfc4729’, ‘22eb0ac0’, ‘239be575’, ‘23b5c85d’, ‘25d8a9c8’, ‘25ff71a9’, ‘272f95fa’, ‘27a28665’, ‘3618c87e’, ‘3af2c5a8’, ‘3bd67248’, ‘3bdb4ada’, ‘44f52bb0’, ‘48d8fb45’, ‘496994bd’, ‘49d1d64f’, ‘539a4f51’, ‘53b68214’, ‘5582e5ca’, ‘5bd6f4ac’, ‘6150a2bd’, ‘62c24649’, ‘67a3c6ac’, ‘67e8384a’, ‘68b16354’, ‘6d0aefbc’, ‘6f8cd79b’, ‘6fa7a44f’, ‘7447852a’, ‘746b3537’, ‘74dd1130’, ‘7837ac64’, ‘794b24be’, ‘7b7f7511’, ‘7f4411dc’, ‘82819916’, ‘88a62173’, ‘8be77c9e’, ‘8e1813be’, ‘90c28cc7’, ‘9172f3a0’, ‘963e52fc’, ‘97999447’, ‘9dfd6313’, ‘a416b8f3’, ‘a65b410d’, ‘a699fb00’, ‘a85d4709’, ‘a87f7484’, ‘aabf363d’, ‘b1948b0a’, ‘b8cdaf2b’, ‘b94a9452’, ‘ba26e723’, ‘ba97ae07’, ‘bc1d5164’, ‘bd4472b8’, ‘bda2d7a6’, ‘bdad9b1f’, ‘c3f564a4’, ‘c59eb873’, ‘c8f0f002’, ‘c9e6f938’, ‘c9f8e694’, ‘d0f5fe59’, ‘d10ecb37’, ‘d13f3404’, ‘d2abd087’, ‘d4469b4b’, ‘d631b094’, ‘d8c310e9’, ‘d9fac9be’, ‘dc433765’, ‘e26a3af2’, ‘e9afcf9a’, ‘ea786f4a’, ‘ef135b50’, ‘f5b8619d’, ‘f76d97a5’

B.1.2 Prompting Setting. The prompts used in comparison across prompting techniques and Inferential Coherence of LLMs are CoT, L2M, and ToT, as shown in Figure 18. These are detailed in Table 9. In the prompts, parts enclosed in curly brackets indicate where the corresponding Type should be inserted. Demo examples and test input refer to the demonstration examples and test input of the test example provided in the task to be solved. For instance, if the type is CoT Prompt, it consists of the CoT one-shot example, the task's demonstration examples, and test input. Regardless of the prompting method, all prompts are given a one-shot example. CoT solves the task using only the one-shot example, the task's demonstration examples, and test input. On the other hand, LtM and ToT use a decomposing prompt to obtain instructions for solving the problem through a decomposing stage. For LtM, the step-by-step solving prompt is then used to execute the instructions obtained through decomposing one by one. The previously executed instructions and the resulting changed grid are included in this process. For ToT, the decomposing prompt is used to create multiple instruction candidates, and the ToT decomposing vote prompt is used to have the LLM select the most promising instruction candidate. The selected instructions are then processed through the step-by-step solving prompt to generate multiple candidate results for each instruction. The ToT step-by-step solving vote prompt is then used to select the grid that best reflects the instruction. This process is carried out step-by-step for all instructions.

Sample Task

If the input grids are:

```
[[0, 3, 0, 0, 0, 0],
 [0, 3, 0, 2, 0, 0],
 [0, 0, 0, 2, 0, 0],
 [0, 8, 0, 0, 2, 2],
 [0, 0, 0, 0, 2, 2],
 [6, 6, 6, 0, 0, 0]]
```

then the correct output grids are:

```
[[0, 0, 0, 0, 3, 0],
 [0, 0, 0, 0, 3, 2],
 [0, 0, 0, 0, 0, 2],
 [0, 0, 0, 8, 2, 2],
 [0, 0, 0, 0, 2, 2],
 [0, 0, 0, 6, 6, 6]]
```

{additional examples}

To solve this task, follow the sub-tasks below.
1. Identify objects in the input grid.
2. Try to move each object to the right.
3. Stop when objects touch the right corner or other objects.

Following these steps will lead to the output grid.

Decomposing

If the input grids are:

```
[[0, 0, 0, 0, 0, 0],
 [0, 0, 3, 0, 0, 0],
 [0, 3, 0, 3, 0, 0],
 [0, 0, 3, 0, 3, 0],
 [0, 0, 0, 3, 0, 0],
 [0, 0, 0, 0, 0, 0]]
```

then the correct output grids are:

```
[[0, 0, 0, 0, 0, 0],
 [0, 0, 3, 0, 0, 0],
 [0, 3, 4, 3, 0, 0],
 [0, 0, 3, 4, 3, 0],
 [0, 0, 0, 3, 0, 0],
 [0, 0, 0, 0, 0, 0]]
```

{additional examples}

To solve this task, decompose the task into sub-tasks like below.
1. Identify the places surrounded by "3"s in the input grid.
2. Fill in the places you found with "4".

Following these steps will lead to the output grid.

Target Task

If the input grids are:

```
[[0, 0, 0, 0, 0, 0, 0],
 [0, 5, 8, 5, 0, 0, 0],
 [0, 5, 8, 5, 0, 0, 0],
 [0, 8, 8, 8, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0]]
```

then the correct output grids are:

```
[[0, 0, 0, 0, 0, 0, 0],
 [0, 8, 5, 8, 0, 0, 0],
 [0, 8, 5, 8, 0, 0, 0],
 [0, 5, 5, 5, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0]]
```

{additional examples}

If the input grid are:

```
[[0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 3, 2, 2, 0, 0, 0, 0, 0],
 [0, 3, 3, 2, 0, 0, 0, 0, 0],
 [0, 3, 2, 2, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 6, 6, 6, 0],
 [0, 0, 0, 0, 0, 1, 1, 1, 0],
 [0, 0, 0, 0, 0, 1, 6, 6, 0]]
```

What is the correct output grid?

Grid Visualization

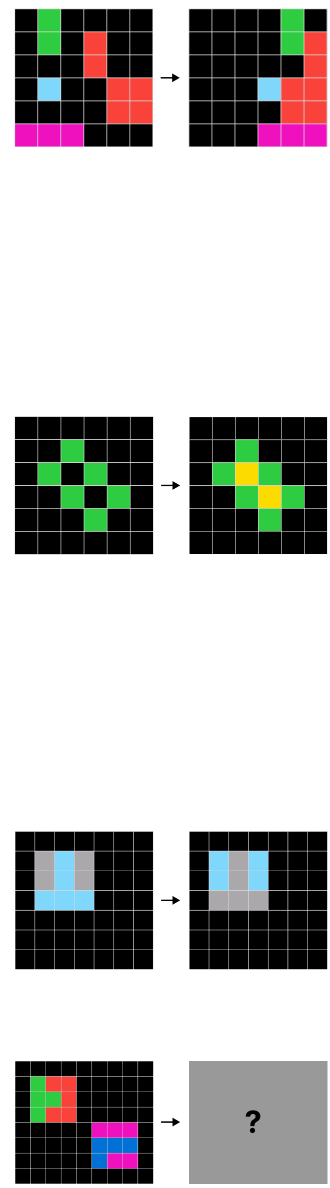


Fig. 17. Three types of prompts are shown on the left. Although all prompts are described as a 2D array of grids, we visualized them on the right for clarity. By default, all three techniques use prompts with two main components: a sample task and a target task. However, LtM and ToT use a different combination of the target task and its decomposition command. This deviation occurs because CoT strictly follows the given sub-task, while LtM and CoT decompose the task on their own.

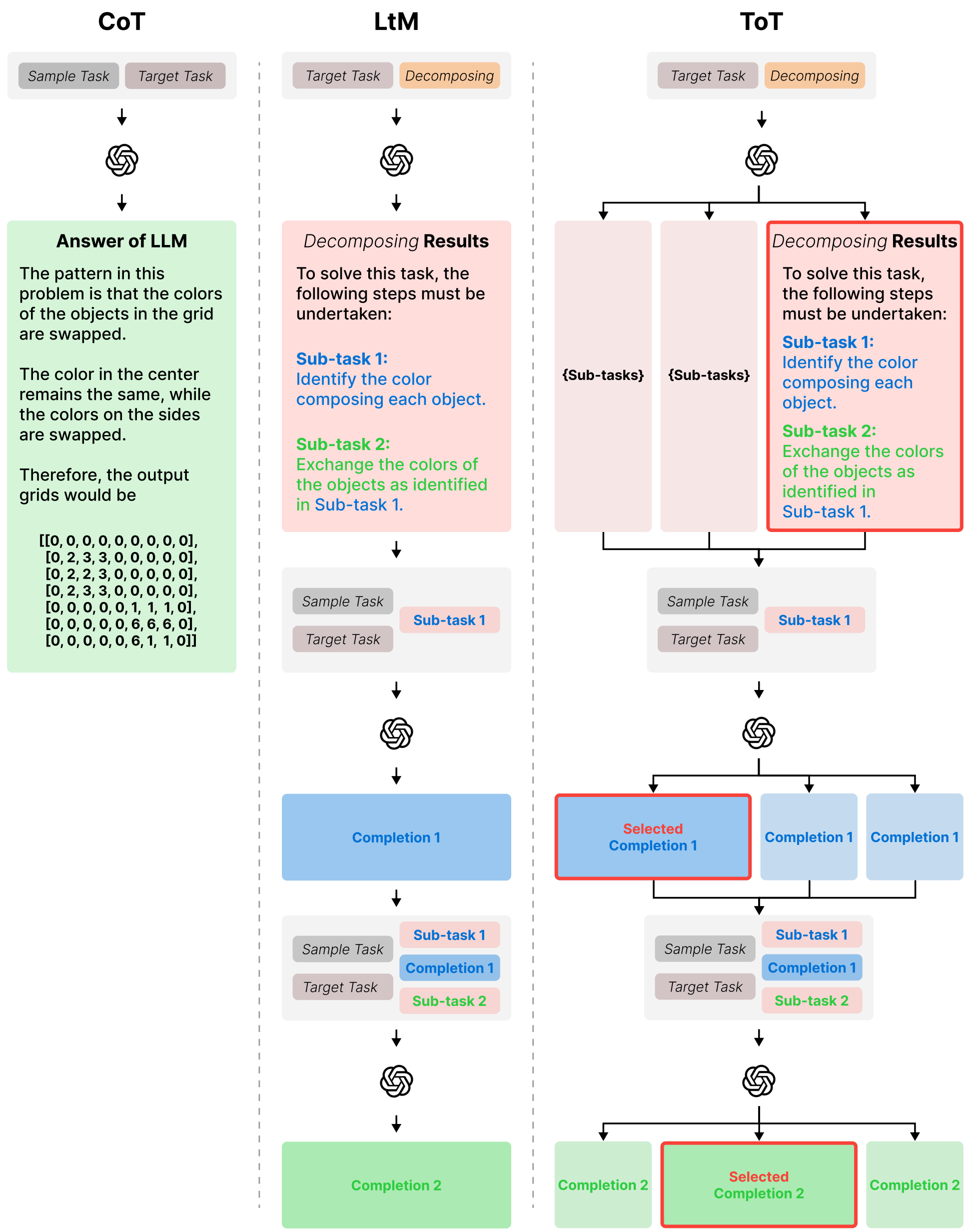


Fig. 18. Grey blocks illustrate prompt sets delivered to the LLM, including the sample task, target task, and LLM's prior responses, as shown in Fig. 17. Green blocks denote the final answer. CoT relies on a single grey block, indicating that the LLM strictly follows the provided sub-tasks. Conversely, LtM and ToT prompt the LLM to generate and address sub-tasks sequentially, represented by decomposed results (red) and intermediate responses (blue). ToT further distinguishes itself from LtM by evaluating multiple suggestions for sub-task handling and selecting the most effective one through a voting mechanism.

Table 9. Prompting Setting Table: The logical coherence experiments employed various prompting techniques including CoT, LtM, and ToT. CoT utilizes the CoT prompt, while LtM uses decomposing and step-by-step solving prompts. ToT incorporates decomposing, ToT decomposing vote, step-by-step solving, and ToT step-by-step solving vote prompts.

Type	**Prompt**
CoT One-Shot Example Data	If input grids are like that [[0, 3, 0, 0, 0, 0], [0, 3, 0, 2, 0, 0], [0, 0, 0, 2, 0, 0], [0, 8, 0, 0, 2, 2], [0, 0, 0, 0, 2, 2], [6, 6, 6, 0, 0, 0]], then these grids change to output grids below [[0, 0, 0, 0, 3, 0], [0, 0, 0, 0, 3, 2], [0, 0, 0, 0, 0, 2], [0, 0, 0, 8, 2, 2], [0, 0, 0, 0, 2, 2], [0, 0, 0, 6, 6, 6]].
CoT One-Shot Example	Do you know ARC problem? It is similar to below. **{CoT One-Shot Example Data}** You can understand the pattern of this problem with the input-output pair in Example 1. In the above case, you can infer as follows. In Example 1, all objects move to the right, and then you can move the object to the right side. Like this concept, ‘object’, ‘count’, ‘color’, ‘move’, ‘row’, ‘column’, etc., help you to understand the patterns of the problem and solve it. Below is another pattern to solve. So you can understand the pattern with several examples and apply the problem’s input to get the correct output.
CoT Prompt	**{CoT One-Shot Example}** **{Demo Examples}** If input grids are like that **{Test Input}** then output grids?

Type	**Prompt**
Decomposing One-Shot Example-Demo Examples	Example 1 If input grids are like that [[0, 0, 0, 0, 0, 0], [0, 0, 3, 0, 0, 0], [0, 3, 0, 3, 0, 0], [0, 0, 3, 0, 3, 0], [0, 0, 0, 3, 0, 0], [0, 0, 0, 0, 0, 0]] then these grids change to output grids below [[0, 0, 0, 0, 0, 0], [0, 0, 3, 0, 0, 0], [0, 3, 4, 3, 0, 0], [0, 0, 3, 4, 3, 0], [0, 0, 0, 3, 0, 0], [0, 0, 0, 0, 0, 0]]. Example 2 If input grids are like that [[0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 3, 0, 3, 0, 0, 0, 0, 0], [0, 0, 0, 3, 0, 3, 0, 0, 0, 0], [0, 0, 3, 0, 0, 0, 3, 0, 0, 0], [0, 0, 0, 0, 0, 3, 0, 3, 0, 0], [0, 0, 0, 3, 0, 3, 3, 0, 0, 0], [0, 0, 3, 3, 3, 0, 0, 0, 0, 0], [0, 0, 0, 3, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0]] then these grids change to output grids below [[0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 3, 0, 3, 0, 0, 0, 0, 0], [0, 0, 0, 3, 0, 3, 0, 0, 0, 0], [0, 0, 3, 0, 0, 0, 3, 0, 0, 0], [0, 0, 0, 0, 0, 3, 4, 3, 0, 0], [0, 0, 0, 3, 0, 3, 3, 0, 0, 0], [0, 0, 3, 3, 3, 0, 0, 0, 0, 0], [0, 0, 0, 3, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0]].
Decomposing One-Shot Example-Test Input	Example-problem If input grids are like that [[0, 0, 0, 0, 0, 3, 0, 0, 0, 0], [0, 0, 0, 0, 3, 0, 0, 0, 0, 0], [0, 3, 3, 0, 3, 3, 0, 3, 0, 0], [3, 0, 0, 3, 0, 0, 3, 0, 3, 0], [0, 0, 0, 3, 0, 0, 3, 3, 0, 0], [0, 0, 0, 3, 0, 0, 3, 0, 0, 0], [0, 0, 0, 3, 0, 0, 3, 0, 0, 0], [0, 0, 0, 0, 3, 3, 0, 3, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 3, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0]] then output grids?

Type	Prompt
Decomposing Prompt	Do you know the ARC problem? Each example has the same pattern and the quiz also has the same pattern with examples. So If you understand the pattern of examples, you can solve the quiz. It will help you to analyze the pattern if you decompose the pattern into some steps. I give an example that decomposes patterns into subquestions. **{Decomposing One-Shot Example-Demo Examples}** **{Decomposing One-Shot Example-Test Input}** To solve the quiz, I think we should do something like below Q1: We need to identify the places surrounded by 3s in the input grid of example-quiz. Q2: Fill in the 4 in the location we found through Q1. **{Demo Examples}** **{Test Input}** I want you to answer in the format below Q1: .... Q2: .... QN: .... N is the index of the last question. (The answers to the last question should allow you to generate the output grid for the quiz, and you shouldn't solve the problem yet in this process. You should only create the subquestions for solving the problem.)

Type	Prompt
Step-By-Step Solving Prompt	**{CoT One-Shot Example}** **{Demo Examples}** If input grids are like that **{Test Input}** then output grids? **{Previous Instructions}** **{Previous Changed Grid}** **{Current Instruction}**
ToT Decomposing Vote Prompt	First, consider how to solve the problem below. **{Demo Examples}** **{Test Input}** Then, given instruction and several choices, decide which choice is most promising. Analyze each choice in detail, then conclude in the last line "The best choice is s", where s is the integer ID of the choice.
ToT Step-By-Step Solving Vote Prompt	First, consider how to solve the problem below. **{Demo Examples}** **{Test Input}** Then, given a question and some answers, you need to choose which answer is the best answer to the question. Analyze each choice in detail and then conclude on the last line, "The best answer is ‘s’," where s is the integer ID of the answer. **{Previous Instructions}** **{Previous Changed Grid}** **{Current Instruction}** **{Current Changed Grid}**

B.2 Compositionality

In the compositionality study, we conducted two experiments: the LLMs DSL understanding experiment to assess how well LLMs comprehend the provided DSLs, and an experiment to evaluate LLMs' compositionality ability. The LLMs DSL understanding experiment measures how accurately LLMs can generate the correct DSLs when given the answer for a task. The compositionality ability experiment examines whether LLMs can correctly select and use the necessary DSLs from those provided for problem-solving. Both experiments used the same set of tasks. Detailed information about the task IDs can be found in Table 11, while the specific Prompt details are available in Table 13 and Table 14.

B.2.1 Task ID List.

Table 11. Task ID list for the compositionality experiment comprising 158 tasks. From the total 800 ARC tasks, we selected only those problems where the input and output grid sizes were identical and could be solved within 10 steps using the given DSL for our experiment.

Experiment	Task ID List
LLMs DSL understanding & Compositionality of LLMs	'025d127b', '05f2a901', '08ed6ac7', '0ca9ddb6', '0d3d703e', '11852cab', '150deff5', '1b60fb0c', '1caeab9d', '1e0a9b12', '1f642eb9', '1f876c06', '2204b7a8', '22168020', '22233c11', '228f6490', '22eb0ac0', '253bf280', '25d487eb', '25d8a9c8', '25ff71a9', '29c11459', '2bee17df', '2c608aff', '2dd70a9a', '3345333e', '3618c87e', '36fdfd69', '3906de3d', '3aa6fb7a', '3bd67248', '3c9b0459', '3e980e27', '3eda0437', '40853293', '42a50994', '444801d8', '44d8ac46', '4938f0c2', '496994bd', '50846271', '508bd3b6', '50cb2852', '5168d44c', '54d82841', '5582e5ca', '56dc2b01', '56ff96f3', '5c0a986e', '60b61512', '6150a2bd', '623ea044', '63613498', '67385a82', '673ef223', '67a3c6ac', '67a423a3', '6855a6e4', '68b16354', '694f12f3', '6c434453', '6cf79266', '6d58a25d', '6d75e8bb', '6e02f1e3', '6e19193c', '6e82a1ae', '74dd1130', '760b3cac', '776ffc46', '794b24be', '7ddcd7ec', '7e0986d6', '7f4411dc', '810b9b61', '855e0971', '88a10436', '890034e9', '8d510a79', '90f3ed37', '928ad970', '93b581b8', '941d9a10', '952a094c', '9565186b', '99fa7670', '9dfd6313', 'a2fd1cf0', 'a3df8b1e', 'a48eeaf7', 'a5313dff', 'a61f2674', 'a65b410d', 'a699fb00', 'a78176bb', 'a79310a0', 'a85d4709', 'a9f96cdd', 'aabf363d', 'ae3edfdc', 'aedd82e4', 'af902bf9', 'b1948b0a', 'b230c067', 'b2862040', 'b548a754', 'b7249182', 'b8cdaf2b', 'ba97ae07', 'bb43febb', 'bda2d7a6', 'bdad9b1f', 'c0f76784', 'c8f0f002', 'cbded52d', 'ce22a75a', 'ce9e57f2', 'd037b0a7', 'd07ae81c', 'd23f8c26', 'd2abd087', 'd406998b', 'd43fd935', 'd4a91cb9', 'd4f3cd78', 'd511f180', 'd5d6de2d', 'd6ad076f', 'd89b689b', 'd8c310e9', 'd90796e8', 'd9f24cd1', 'dc433765', 'ddf7fa4f', 'ded97339', 'e40b9e2f', 'e48d4e1a', 'e5062a87', 'e509e548', 'e73095fd', 'e8dc4411', 'e9614598', 'e9afcf9a', 'ea32f347', 'ea786f4a', 'ec883f72', 'ed36ccf7', 'ef135b50', 'f25ffba3', 'f76d97a5', 'f8a8fe49', 'fcc82909', '8f2ea7aa', '5521c0d9', '32597951', '98cf29f8', '0e206a2e', 'a1570a43'

B.2.2 Types of DSLs used. Each DSL was implemented as a Python function. As shown in Table 12, there are three types of DSLs using three parameter types. Color Change DSLs accept parameters such as Coordinate and Object. Coordinate-based Color Change DSLs include pixel color, X line, horizontal line, vertical line, and diagonal line. For Object parameters, only the obj color DSL exists. Transformation DSLs use Object and Grid parameters. Object-based transformations include rotate left obj, rotate right obj, horizontal flip obj, vertical flip obj, and movement operations (move left, right, up, down). Grid-based transformations include rotate left state, rotate right state, horizontal flip, and vertical flip. Lastly, the Complete DSL exists independently of parameters, indicating task completion before reaching 10 steps. For tasks solved in exactly 10 steps, the Complete DSL is unnecessary.

Table 12. DSL list: The DSL used in the compositionality experiment is categorized based on the type of target and the kind of functionality. The targets in the DSL are categorized into three types: Coordinate, Object, and Grid. The functions of the DSL include changing the color of a target (Color Change), moving a target (Transformation), and indicating the completion of the task 10 steps earlier (Complete).

	Coordinate	**Object**	**Grid**
Color Change	pixel color, X line, horizontal line, vertical line, diagonal line	obj color	X
Transformation	X	rotate left obj, rotate right obj, horizontal flip obj, vertical flip obj, move left, move right, move up, move down	rotate left state, rotate right state, horizontal flip, vertical flip
Complete		Complete	

B.2.3 Prompt Contents for LLMs with DSL Codes and Comments . In the two experiments measuring compositionality and LLM's DSL understanding, we identified a set of 10 tasks that collectively required the use of all 15 DSLs at least once. This set was used to determine the optimal prompt for explaining DSLs to LLMs. We conducted experiments with four prompt variants: no DSL information, DSL code only, DSL comments only, and both DSL code and comments. The LLMs DSL understanding experiment was performed for these 10 tasks across all four prompt compositions. Results indicated that providing both code and comments yielded optimal performance. Consequently, for both the LLMs DSL understanding and compositionality of LLMs experiments, we employed prompts containing both DSL code and comments. Table 13 illustrates the prompt content where both code and comments were provided to the LLM.

Table 13. Prompts of DSL Function Codes and Comments

<table>
<tr><th>Type</th><th>Prompt</th></tr>
<tr><td>Rotate Left State</td><td><pre>
rotate_left_state function is a counterclockwise rotation about the given state.
This function rotates a square grid (NxN) counterclockwise by 90 degrees.
Parameters:
- state: A 2D list representing the current grid state.
Returns:
- A new 2D list representing the grid after the counterclockwise rotation.

def rotate_left_state(state):
    N = len(state)
    rotated_state = copy.deepcopy(state)
    if N == len(state[0]):
        temp_state = copy.deepcopy(state)
        for x in range(N):
            for y in range(N):
                rotated_state[N-1-y][x] = state[x][y]
    return rotated_state
</pre></td></tr>
<tr><td>Rotate Right State</td><td><pre>
rotate_right_state function is a clockwise rotation about the given state.
This function rotates a square grid (NxN) clockwise by 90 degrees.
Parameters:
- state: A 2D list representing the current grid state.
Returns:
- A new 2D list representing the grid after the clockwise rotation.

def rotate_right_state(state):
    N = len(state)
    rotated_state = copy.deepcopy(state)
    if N == len(state[0]):
        for x in range(N):
            for y in range(N):
                rotated_state[y][N-1-x] = state[x][y]
    return rotated_state
</pre></td></tr>
</table>

Type	Prompt

Vertical Flip	# vertical_flip function is a flip by x-axis about the given state.
	# This function flips the grid vertically, swapping the top and bottom rows.
	# Parameters:
	# - state: A 2D list representing the current grid state.
	# Returns:
	# - A new 2D list representing the grid after the vertical flip.
	
	def vertical_flip(state):
	    temp_state = copy.deepcopy(state)
	    N = len(state)
	    M = len(state[0])
	        for i in range(N):
	            for j in range(M):
	                temp_state[N-1-i][j] = state[i][j]
	    return temp_state

Horizontal Flip	# horizontal_flip function is a flip by y-axis about the given state.
	# This function flips the grid horizontally, swapping the left and right columns.
	# Parameters:
	# - state: A 2D list representing the current grid state.
	# Returns:
	# - A new 2D list representing the grid after the horizontal flip.
	
	def horizontal_flip(state):
	    N = len(state)
	    M = len(state[0])
	    flipped_state = copy.deepcopy(state)
	    for i in range(N):
	        for j in range(M // 2):
	        flipped_state[i][j], flipped_state[i][M-1-j] = state[i][M-1-j], state[i][j]
	    return flipped_state

Type	Prompt

Move Right	# move_right function moves all pixels in the selected object to the right by one column.
	# Parameters:
	# - state: A 2D list representing the current grid state.
	# - object: A list of lists where each inner list contains the coordinates [x, y] of a pixel to move.
	# Returns:
	# - A new 2D list representing the grid after the object is moved to the right.
	
	def move_right(state, object):
	    move_state = copy.deepcopy(state)
	    new_obj = []
	
	    for x, y in object:
	        move_state[x][y] = 0
	
	    for x, y in object:
	        new_x, new_y = x, y + 1
	        if 0 <= new_x < len(state) and 0 <= new_y < len(state[0]):
	            move_state[new_x][new_y] = state[x][y]
	            new_obj.append([new_x, new_y])
	
	    return move_state

Type	**Prompt**

Move Left	# move_left function moves all pixels in the selected object to the left by one column.
	# Parameters:
	# - state: A 2D list representing the current grid state.
	# - object: A list of lists where each inner list contains the coordinates [x, y] of a pixel to move.
	# Returns:
	# - A new 2D list representing the grid after the object is moved to the left.
	
	def move_left(state, object):
	    move_state = copy.deepcopy(state)
	    new_obj = []
	
	    for x, y in object:
	        move_state[x][y] = 0
	
	    for x, y in object:
	        new_x, new_y = x, y - 1
	        if 0 <= new_x < len(state) and 0 <= new_y < len(state[0]):
	            move_state[new_x][new_y] = state[x][y]
	            new_obj.append([new_x, new_y])
	
	    return move_state

Type	Prompt

Move Up	# move_up function moves all pixels in the selected object up by one row.
	# Parameters:
	# - state: A 2D list representing the current grid state.
	# - object: A list of lists where each inner list contains the coordinates [x, y] of a pixel to move.
	# Returns:
	# - A new 2D list representing the grid after the object is moved up.
	
	def move_up(state, object):
	    move_state = copy.deepcopy(state)
	    new_obj = []
	
	    for x, y in object:
	        move_state[x][y] = 0
	
	    for x, y in object:
	        new_x, new_y = x-1, y
	        if 0 <= new_x < len(state) and 0 <= new_y < len(state[0]):
	            move_state[new_x][new_y] = state[x][y]
	            new_obj.append([new_x, new_y])
	
	    return move_state

Type	Prompt

Move Down	# move_down function moves all pixels in the selected object down by one row.
	# Parameters:
	# - state: A 2D list representing the current grid state.
	# - object: A list of lists where each inner list contains the coordinates [x, y] of a pixel to move.
	# Returns:
	# - A new 2D list representing the grid after the object is moved down.
	
	def move_down(state, object):
	    move_state = copy.deepcopy(state) new_obj = []
	
	    for x, y in object:
	        move_state[x][y] = 0
	
	    for x, y in object:
	        new_x, new_y = x+1, y
	        if 0 <= new_x < len(state) and 0 <= new_y < len(state[0]):
	            move_state[new_x][new_y] = state[x][y]
	            new_obj.append([new_x, new_y])
	
	    return move_state

Type	Prompt

Rotate Right Object	# rotate_right_obj function makes a clockwise rotation about the given object.
	# This function rotates the selected object within the grid 90 degrees clockwise around its center.
	# Parameters:
	# - state: A 2D list representing the current grid state.
	# - object: A list of lists where each inner list contains the coordinates [x, y] of a pixel in the object.
	# Returns:
	# - A new 2D list representing the grid after the object is rotated clockwise.
	
	def rotate_right_obj(state, object):
	    rotate_state = copy.deepcopy(state)
	    new_obj = []
	    max_x = max(x for x,_ in object)
	    min_x = min(x for x,_ in object)
	    max_y = max(y for_, y in object)
	    min_y = min(y for_, y in object)
	    fixed_x = (max_x + min_x) // 2
	    fixed_y = (max_y + min_y) // 2
	
	    for x, y in object:
	        rotate_state[x][y] = 0
	
	    for x, y in object:
	        moved_x = y - fixed_y + fixed_x
	        moved_y = -x + fixed_x + fixed_y
	        if 0 <= moved_x < len(state) and 0 <= moved_y < len(state[0]):
	            rotate_state[moved_x][moved_y] = state[x][y]
	            new_obj.append([moved_x, moved_y])
	
	    return rotate_state

Type	Prompt

Rotate Left Object	# rotate_left_obj function makes a counterclockwise rotation about the given object.
	# This function rotates the selected object within the grid 90 degrees counterclockwise around its center.
	# Parameters:
	# - state: A 2D list representing the current grid state.
	# - object: A list of lists where each inner list contains the coordinates [x, y] of a pixel in the object.
	# Returns:
	# - A new 2D list representing the grid after the object is rotated counterclockwise.
	
	def rotate_left_obj(state, object):
	    rotate_state = copy.deepcopy(state)
	    new_obj = []
	    max_x = max(x for x,_ in object)
	    min_x = min(x for x,_ in object)
	    max_y = max(y for_, y in object)
	    min_y = min(y for_, y in object)
	    fixed_x = (max_x + min_x) // 2
	    fixed_y = (max_y + min_y) // 2
	
	    for x, y in object:
	        rotate_state[x][y] = 0
	
	    for x, y in object:
	        moved_x = -y + fixed_y + fixed_x
	        moved_y = x - fixed_x + fixed_y
	        if 0 <= moved_x < len(state) and 0 <= moved_y < len(state[0]):
	            rotate_state[moved_x][moved_y] = state[x][y]
	            new_obj.append([moved_x, moved_y])
	
	    return rotate_state

Type	Prompt

Vertical Flip Object	# vertical_flip_obj function makes a vertical flip of the selected object.
	# This function flips the selected object within the grid vertically.
	# Parameters:
	# - state: A 2D list representing the current grid state.
	# - object: A list of lists where each inner list contains the coordinates [x, y] of a pixel in the object.
	# Returns:
	# - A new 2D list representing the grid after the object is flipped vertically.
	
	def vertical_flip_obj(state, object):
	    flip_state = copy.deepcopy(state)
	    new_obj = []
	    max_x = max(x for x,_ in object)
	    min_x = min(x for x,_ in object)
	
	    for x, y in object:
	        flip_state[x][y] = 0
	
	    for x, y in object: flip_state[max_x + min_x - x][y] = state[x][y]
	        new_obj.append([max_x + min_x - x, y])
	
	    return flip_state

Type	**Prompt**

Horizontal Flip Object	# horizontal_flip_obj function makes a horizontal flip of the selected object.
	# This function flips the selected object within the grid horizontally.
	# Parameters:
	# - state: A 2D list representing the current grid state.
	# - object: A list of lists where each inner list contains the coordinates [x, y] of a pixel in the object.
	# Returns:
	# - A new 2D list representing the grid after the object is flipped horizontally.
	
	def horizontal_flip_obj(state, object):
	    flip_state = copy.deepcopy(state)
	    new_obj = []
	    max_y = max(y for_, y in object)
	    min_y = min(y for_, y in object)
	
	    for x, y in object:
	        flip_state[x][y] = 0
	
	    for x, y in object:
	        flip_state[x][max_y + min_y - y] = state[x][y]
	        new_obj.append([x, max_y + min_y - y])
	
	    return flip_state

<table>
<tr><th>Type</th><th>Prompt</th></tr>
<tr><td>X Line</td><td># X_line function makes a diagonal X-line in all directions from a given pixel until they reach the end of the grid.
Parameters:
- state: A 2D list representing the current grid state.
- r: The row index of the starting pixel.
- c: The column index of the starting pixel.
- color: The color to be used for the X-line.
Returns:
- A new 2D list representing the grid after the X-line is drawn.

def X_line(state, r, c, color):
    X_state = copy.deepcopy(state)
    x_move = [-1, 1]
    y_move = [-1, 1]

    for i in x_move:
        for j in y_move:
            moved_x, moved_y = r + i, c + j
            while 0 <= moved_x < len(state) and 0 <= moved_y < len(state[0]):
                X_state[moved_x][moved_y] = color
                moved_x += i
                moved_y += j

    return X_state</td></tr>
</table>

Type	Prompt

Horizontal Line	# horizontal_line function draws a horizontal line between two pixels if they are on the same row.
	# Parameters:
	# - state: A 2D list representing the current grid state.
	# - r1: The row index of the first pixel.
	# - c1: The column index of the first pixel.
	# - r2: The row index of the second pixel.
	# - c2: The column index of the second pixel.
	# - color: The color to be used for the line.
	# Returns:
	# - A new 2D list representing the grid after the horizontal line is drawn.
	
	def horizontal_line(state, r1, c1, r2, c2, color):
	    line_state = copy.deepcopy(state)
	    if r1 == r2:
	        if c1 < c2:
	            if c2 < len(state[0]):
	                for i in range(c1+1, c2):
	                    line_state[r1][i] = color
	        else:
	            if c1 < len(state[0]):
	                for i in range(c2+1, c1):
	                    line_state[r1][i] = color
	    return line_state

<table>
<tr><th>Type</th><th>Prompt</th></tr>
<tr><td>Vertical Line</td><td><pre>
vertical_line function draws a vertical line between two pixels if they are in the same column.
Parameters:
- state: A 2D list representing the current grid state.
- r1: The row index of the first pixel.
- c1: The column index of the first pixel.
- r2: The row index of the second pixel.
- c2: The column index of the second pixel.
- color: The color to be used for the line.
Returns:
- A new 2D list representing the grid after the vertical line is drawn.

def vertical_line(state, r1, c1, r2, c2, color):
    line_state = copy.deepcopy(state)
    if c1 == c2:
        if r1 < r2:
            if r2 < len(state):
                for i in range(r1+1, r2):
                    line_state[i][c1] = color
        else:
            if r1 < len(state):
                for i in range(r2+1, r1):
                    line_state[i][c1] = color
    return line_state
</pre></td></tr>
</table>

<table>
<tr><th>Type</th><th>Prompt</th></tr>
<tr><td>Diagonal Line</td><td><pre>
diagonal_line function draws a diagonal line between two pixels if they form a 45-degree angle.
Parameters:
- state: A 2D list representing the current grid state.
- r1: The row index of the first pixel.
- c1: The column index of the first pixel.
- r2: The row index of the second pixel.
- c2: The column index of the second pixel.
- color: The color to be used for the line.
Returns:
- A new 2D list representing the grid after the diagonal line is drawn.

def diagonal_line(state, r1, c1, r2, c2, color):
    line_state = copy.deepcopy(state)
    if abs(r1 - r2) == abs(c1 - c2):
        dr = 1 if r2 > r1 else -1
        dc = 1 if c2 > c1 else -1
        r, c = r1 + dr, c1 + dc
        while r != r2 and c != c2:
            line_state[r][c] = color
            r += dr
            c += dc
    return line_state
</pre></td></tr>
<tr><td>Object Color</td><td><pre>
obj_color function changes the color of the selected object.
Parameters:
- state: A 2D list representing the current grid state.
- object: A list of lists where each inner list contains the coordinates [x, y] of a pixel in the object.
- color: The new color to be applied to the object.
Returns:
- A new 2D list representing the grid after the object's color is changed.

def obj_color(state, object, color):
    color_state = copy.deepcopy(state)
    for x, y in object:
        color_state[x][y] = color
    return color_state
</pre></td></tr>
</table>

<table>
<tr><th>Type</th><th>Prompt</th></tr>
<tr><td>Pixel Color</td><td><pre># pixel_color function changes the color of the selected pixel.
Parameters:
- state: A 2D list representing the current grid state.
- r: The row index of the pixel to change.
- c: The column index of the pixel to change.
- color: The new color to be applied to the pixel.
Returns:
- A new 2D list representing the grid after the pixel’s color is changed.

def pixel_color(state, r, c, color):
    temp_state = copy.deepcopy(state)
    temp_state[r][c] = color
    return temp_state</pre></td></tr>
<tr><td>Complete</td><td><pre># complete function returns the current state as the final answer of the quiz.
Parameters:
- state: A 2D list representing the current grid state.
Returns:
- The same 2D list representing the final grid state.

def complete(state):
    return state</pre></td></tr>
</table>

B.2.4 Prompt of Compositionality Experiment. Both the LLMs DSL understanding and compositionality of LLMs experiments utilized the prompt structure outlined in Table 14. The Introduction ARC Prompt provides a comprehensive overview of ARC, while the DSL Usage Example Prompt illustrates DSL application. The DSL prompt, comprising the prompts of DSL function codes and comments from Table 13 and the DSL usage example prompt, offers a comprehensive DSL explanation. The task prompt includes demonstration examples, test input, object information (coordinates of objects obtained through PnP in dictionary format), and output format guidelines. In the case of the LLMs DSL understanding prompt, unlike the task prompt, the DSLs path for the task is provided. The CoT prompt, included the introduction ARC prompt and DSL prompt. In the case of the LLMs DSL understanding experiment, the LLMs DSL understanding prompt was used, while in the case of the compositionality of LLMs experiment, the task prompt was used. In the compositionality experiments, the CoT prompt was utilized.

Table 14. Prompt table: Composition of prompt contents used in the compositionality experiments.

Type	Prompt
Introduction ARC Prompt	Do you know ARC problem? ARC is a quiz, and if we can solve this problem, we can understand and utilize several concepts such as "object", "count", "color", "move", "row", "column", etc. ARC problems give you some examples to understand these patterns. You can understand the pattern below with several examples and then apply the quiz's input to get the correct output.
DSL Usage Example Prompt	In this example grid, [[0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 5, 5, 0], [0, 5, 5, 0, 0, 0, 0, 5, 5, 0], [0, 0, 5, 5, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 5], [0, 0, 0, 0, 0, 5, 5, 0, 0, 5], [0, 5, 0, 0, 0, 0, 0, 0, 0, 5], [0, 5, 0, 0, 5, 0, 0, 0, 0, 0], [0, 0, 0, 5, 5, 0, 0, 0, 0, 0]] there are 6 objects Object1: [[1, 7], [1, 8], [2, 7], [2, 8]] Object2: [[2, 1], [2, 2], [3, 2], [3, 3]] Object3: [[5, 9], [6, 9], [7, 9]] Object4: [[6, 5], [6, 6]] Object5: [[7, 1], [8, 1]] Object6: [[8, 4], [9, 3], [9, 4]] If you apply "rotate_right_obj(state, object2)", the result becomes [[0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 5, 0, 0, 0, 0, 5, 5, 0], [0, 5, 5, 0, 0, 0, 0, 5, 5, 0], [0, 5, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 5], [0, 0, 0, 0, 0, 5, 5, 0, 0, 5], [0, 5, 0, 0, 0, 0, 0, 0, 0, 5], [0, 5, 0, 0, 5, 0, 0, 0, 0, 0], [0, 0, 0, 5, 5, 0, 0, 0, 0, 0]]

Type	Prompt
DSL Prompt	**{Prompts of DSL Function Codes and Comments}** Arguments for the DSLs are mainly "state" and "object", but some require "color", "row", "column", etc. "state" is the current state of the grid, which is the entire grid. **"object"** is the list of coordinates of the object; there may be multiple objects in the grid, but no DSL requires multiple objects. **"color"** is the color of the pixel in the grid, which is a number between 0 and 9. **"row"** and **"column"** are the coordinate numbers of a pixel in the grid. You can choose from here and apply the DSL to solve the problem. You must input the appropriate arguments to the DSL, or it will not work. **{DSL Usage Example Prompt}** Please choose the DSL from the list above and provide the proper arguments to solve the problem.

Task Prompt	**{Demo Examples}** & **{Test Input}** & **{Object Info}**
	
	You must solve the given quiz within 10 steps! Select one DSL with proper arguments in each step. If you think the current state is correct, you can select the "complete" DSL.
	
	I want you to answer in the format below.
	
	The output should be in the following JSON format:
	{
	‘step’: "(current_step)",
	‘dsl’: "(dsl with the arguments for the DSL)",
	‘description’: "(why you chose this DSL?)"
	}

Type	Prompt

LLMs DSL Understanding Prompt	**{Demo Examples}** & **{Test Input}** & **{Object Info}** & **{DSLs Path}**
	
	Generate the grid step by step by applying the DSLs in the 'dsls_path' to the input grid.
	
	The output should be in the following JSON format:
	{
	'step': 'Step number'
	'dsl': 'DSL used from the dsls_path to make the final output grid'
	'grid': 'Final output grid'
	'objects': 'Objects in the final output grid'
	}

CoT Prompt	**{Introduction ARC Prompt}** **{DSL Prompt}** **{LLMs DSL Understanding Prompt}** # It is used when conducting LLMs DSL understanding experiment. **{Task Prompt}** # It is used when conducting compositionality of LLMs experiment.

B.3 Productivity

In the productivity experiment, we aimed to augment the task's demonstration example pairs using the Inverse Transformation Prompt (ITP). The ITP consists of a category prompt, which contains a description of the category, task examples, and the target output to be augmented. The detailed structure of the category prompt is provided in Table 15, and the structure of the ITP is outlined in Table 16.

Table 15. Category Table: This contains prompts explaining the 16 types of Concept ARC. In Table 16, the category prompt is filled with the appropriate prompt corresponding to the category of the task to be generated.

Category	**Prompt**
AboveBelow	Focus on the horizontal criteria, you may have to modify some regein by that line. such as removing, moving, filling region by color. element. See the provided example to how to modify.
Center	Fix the array issue by addressing the center, potentially moving or removing the central element. See the provided example for clarity.
CleanUp	Distort the shapes in areas where they are polygonal or completely filled, adding noise or disturbances to disrupt the complete shapes.
CompleteShape	Distort the perfectly shaped objects identified in the input image. Introduce noise to these identified objects to easily generate diverse outputs.
Copy	Delete one identical object from the output. Refer to the example to identify which one to remove. Consider deleting the object located in a position-indicating space.
Count	Create an input image based on the provided count-related problem. Focus on details like object or color count, as shown in the example.
ExtendToBoundary	In the input image, find lines connected to boundaries with different colors. Transform these lines into a different shape. The example illustrates how to make this transformation.
ExtractObjects	Generate an output image with objects from the given input. Refer to examples for guidance. Hint: Extract objects when inferring input from output.
FilledNotFilled	When inferring the input from the output, focus on situations where the inner part of an object contains empty space or another object. Examples provide guidance for creating the output image.
HorizontalVertical	Focus on horizontal and vertical relations, representing them with colors or preserving one direction while eliminating the other. Examples illustrate the approach.

Category	Prompt
InsideOutside	Address the inside-outside relationship, either by selecting items inside or outside in the input or determining quantities. Use the boundary as a reference. Examples offer guidance.
MoveToBoundary	Objects in the input may be shifted to one side, and in the output, they are displaced either horizontally or vertically. Infer the direction from examples and choose the displacement freely.
Order	This is about randomly rearranging initially ordered objects while representing their original positions through a specific rule. Examine the examples to understand how to achieve this.
SameDifferent	You'll notice that only specific-shaped objects are extracted in the input image. Create additional objects in the zero-represented space. Examples provide guidance on how to proceed.
TopBottom2D	Objects are in a 2D space. Check changes in the top and bottom. The input may have shifted or require removing top/bottom indicators. Look at examples for specifics.
TopBottom3D	Objects are in 3D space. Consider front-back relationships; bring objects forward or move them backward in the input. Look at examples for specific methods.

Table 16. ITP: Composition of prompt contents used in the productivity experiments. ITP consists of category prompt, task examples and target output.

Type	Prompt
ITP	Try solving the ARC problem and do not say sensitive word. : generate the output accordingly. I will give you Hint, **{Category Prompt}** Here are some examples to help you. **{Tasks Examples}** **{Target Output}** Provide two dimensional array that are not identical to the input array or a direct copy of the example. Each element is an integer between 0 and 9. Would you give me 2 answer? No need to explain how you solved.