From Generation to Selection: Findings of Converting Analogical Problem-Solving into Multiple-Choice Questions

Donghyeon Shin[1]*, **Seungpil Lee**[1]*, **Klea Lena Kovačec**[1], **Sundong Kim**[1]†
[1]Gwangju Institute of Science and Technology
{shindong97411,iamseungpil,klealk8,sdkim0211}@gmail.com

Abstract

As artificial intelligence reasoning abilities gain prominence, generating reliable benchmarks becomes crucial. The Abstract and Reasoning Corpus (ARC) offers challenging problems yet unsolved by AI. While ARC effectively assesses reasoning, its generation-based evaluation overlooks other assessment aspects. Bloom's Taxonomy suggests evaluating six cognitive stages: *Remember*, *Understand*, *Apply*, *Analyze*, *Evaluate*, and *Create*. To extend ARC's focus beyond the *Create* stage, we developed MC-LARC, a multiple-choice format suitable for assessing stages like *Understand* and *Apply* in Large Language Models (LLMs). Our evaluation of ChatGPT4V's analogical reasoning using MC-LARC confirmed that this format supports LLMs' reasoning capabilities and facilitates evidence analysis. However, we observed LLMs using shortcuts in MC-LARC tasks. To address this, we propose a self-feedback framework where LLMs identify issues and generate improved options. MC-LARC is available at https://mc-larc.github.io/.

1 Introduction

Research on artificial intelligence with reasoning capabilities is attracting attention, leading to the proposal of benchmarks to measure such abilities. The Abstraction and Reasoning Corpus (ARC) is one such benchmark designed to evaluate reasoning abilities. Each ARC task consists of 2–5 examples where both input and output are provided, along with one task where only the input is given. The goal is to infer the rule from the examples and deduce the answer to the task. The input and output grids in ARC can range from a minimum 1×1 grid to a maximum 30×30 grid, with each grid filled with up to 10 different colors. Unlike existing reasoning benchmarks, ARC's strength lies in its specialization in evaluating reasoning abilities alone by reducing the amount of prior knowledge and data required to solve the tasks.

However, ARC has limitations in that it is an overly difficult benchmark requiring multiple stages of reasoning to solve. According to Bloom's Taxonomy (Anderson et al., 2001), proposed in traditional educational theory, evaluation consists of the following six stages: *Remember*, *Understand*, *Apply*, *Analyze*, *Evaluate*, and *Create*. In this taxonomy, ARC assesses creation, which encompasses all prior levels of cognitive processes, making it difficult to pinpoint which specific stage may be problematic when a solution is not reached. Even if the logical reasoning process is correct, the entire response is marked as wrong if there is a slight error in the generated grid. This issue is also found in derived datasets with reduced difficulty, such as Mini-ARC (Kim et al., 2022) and 1D-ARC (Xu et al., 2023). Although these datasets changed grid sizes or reduced 2D arrays to 1D arrays, it remains difficult to identify which part of the model's reasoning process is flawed when the task is not solved due to the evaluation format that includes creation. Therefore, a new evaluation method is needed to identify which step of reasoning is problematic in solving ARC.

Therefore, this paper proposes a modified benchmark called MC-LARC to provide an intermediate step in solving ARC tasks. MC-LARC aims to convert the evaluation format from generation to selection, assessing the areas corresponding to *Understand* and *Apply* in Bloom's Taxonomy. It converts the dataset into a multiple-choice language format by using Large Language Models (LLMs) to generate four alternative options based on the correct answer to ARC tasks. We conducted experiments to investigate the impact of the transformation into multiple-choice form and found the following two points: 1) The accuracy of LLMs on ARC tasks increased from about 10% to 76%. This

*The authors contribute equally.
†Corresponding Author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 6696–6708

modification suggests a narrowing of ARC's assessment scope from multiple cognitive processes to primary comprehension. 2) Evaluating the extent of the inferential abilities of LLMs becomes more clearly feasible. However, it was observed that LLMs used shortcuts while solving MC-LARC, finding the correct answer by considering the form or internal context of the choices to eliminate inappropriate options, rather than utilizing reasoning abilities. To address this issue, we introduce a self-feedback framework that leverages LLMs to improve shortcuts. This method extends beyond previous constraint-based augmentation by incorporating three additional steps: the LLM attempts to solve the multiple-choice questions, articulates the problem situations, and then refines the options, thereby autonomously mitigating shortcuts.

2 Related Works

2.1 Evaluation Methods for LLM Abilities Based on Bloom's Taxonomy

Bloom's Taxonomy (Anderson et al., 2001) provides a hierarchical classification of cognitive skills that educators can use to structure learning objectives, assessments, and activities. The taxonomy categorizes cognitive skills into six levels as illustrated in Figure 1, each representing a different level of complexity and depth of understanding, from the most basic (*Remembering*) to the most advanced (*Creating*).

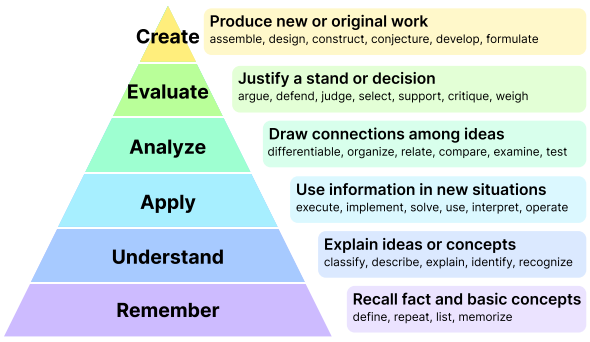

Figure 1: The six cognitive skills in Bloom's Taxonomy. These skills start with basic tasks like recalling facts and understanding concepts at the bottom and progress to creating original work based on a deep understanding of a concept at the top. Image credits: Center for Teaching, Vanderbilt University (Armstrong, 2010).

By utilizing Bloom's Taxonomy, educators and researchers can more effectively design, evaluate, and enhance learning experiences and assessments, ensuring that they address all levels of cognitive skills, from basic recall of information to the creation of new and original work.

Shojaee-Mend et al. (2024) employed Bloom's Taxonomy to assess the cognitive levels of neurophysiology questions answered by large language models, revealing strengths in basic knowledge recall and weaknesses in higher-order reasoning and knowledge integration. Similarly, Joshi et al. (2024) used this taxonomy to analyze the cognitive depth of recommendations made by ChatGPT and Bard for teaching Parallel Coordinate Plots. Human-expert evaluations showed that ChatGPT's suggestions were generally more appropriate and effective across various cognitive stages, while Bard's recommendations were often less reliable. Additionally, the BloomGPT project (Spanos et al., 2024) structured a ChatGPT-powered web application around Bloom's Taxonomy, enhancing students' cognitive and metacognitive learning in an undergraduate history course. Expert evaluations indicated that the application effectively supported diverse cognitive processes.

2.2 Benchmarks for Abstraction Tasks

Abstraction and Reasoning Corpus (ARC) The Abstraction and Reasoning Corpus (ARC) benchmark (Chollet, 2019) was created for the purpose of measuring intelligence in computer systems. This benchmark requires inference based on complex prior knowledge such as arithmetic abilities, geometric understanding, and topological understanding. The goal is to derive common rules from examples and apply them to infer the appropriate output image for a given test input image. Each task provides 2–5 pairs of example input and output images. The original ARC benchmark consists of 400 training set, 400 evaluation set, and 200 test set. Moreover, the ARC benchmark is represented as 2D matrices.

Language-Complete ARC (LARC) The LARC (Acquaviva et al., 2022) dataset consists of 400 ARC training data, each accompanied by 1) a description of the input image and 2) a natural language description of the rules between the input and output images. Both the input description and the output description must be language-complete. Language-complete refers to having sufficient relevant information even when neither input nor output images are provided. In other words, humans should be able to understand the task sufficiently based solely on the description of LARC without

the presence of images. A language-complete ARC is shown in the Refined LARC in Figure 2.

Modified Benchmark with Transformed Evaluation Format Abstract and reasoning tasks often face problems in setting task objectives due to their attempt to measure unclearly defined reasoning abilities. Therefore, there have been previous studies that tried to perform new tasks by modifying or expanding existing tasks. Bongard-LOGO (Nie et al., 2020) is an example of simplifying a complex task. Bongard (Bongard, 1968), one of the Visual Reasoning benchmarks, is a task that expresses the difference between two given abstract image groups as a natural language description. It has long been a notable task as it requires high abstraction and reasoning ability to solve the problem, but it had limitations in analyzing the cause when a specific model could not solve it, as it is a description task requiring natural language processing abilities. To address this, Bongard-LOGO transformed the type of Bongard problem from a description task to a classification task. On the other hand, there are also cases where simple tasks were changed into complex tasks. VQA (Antol et al., 2015) is a task that evaluates how well one can answer when given an image and a question. However, VQA only assesses whether the given image and natural language problem is well understood, making it unsuitable for evaluating reasoning abilities. To overcome this limitation, a modified benchmark, TGIF-QA (Jang et al., 2017), which added questions requiring reasoning about visual images, was proposed. Thus, especially in the field of Visual Reasoning, attempts are being made to establish intermediary results through task transformation.

3 MC-LARC: Generation to Selection

We created MC-LARC through the following two steps: 1) manually refining the existing LARC, and 2) utilizing ChatGPT4 to generate wrong options based on LARC.

Refining Process The original LARC had notable quality issues, as shown in Figure 2. These issues mainly involved 1) inconsistent expressions for the same concept and 2) insufficient details in the provided explanations. For example, the upper part of Figure 2 shows different ways of representing the same concepts, causing confusion. This inconsistency could lead to issues when using language models to augment incorrect options, as the

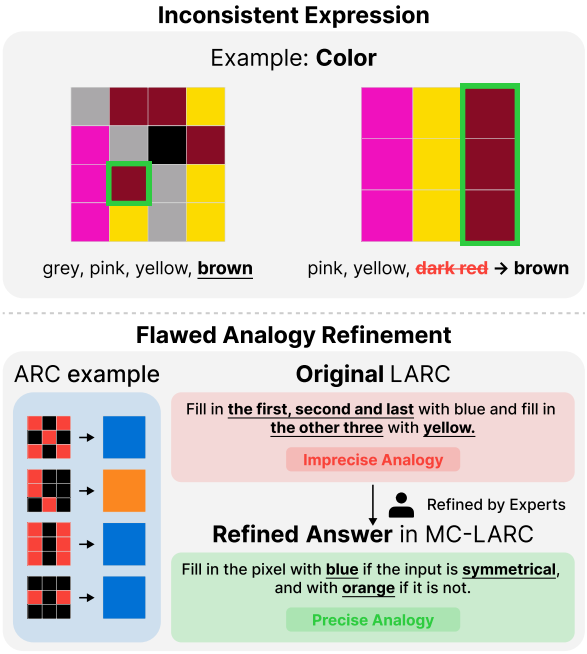

Figure 2: Two main issues of LARC. (Upper) Inconsistent terminology: Varied expressions for identical concepts (e.g., 'brown' described as 'dark red'). (Lower) Insufficient problem-solving information: Original LARC descriptions lack critical details for ARC task completion (e.g., symmetry identification). Expert revisions fill in these missing details.

model might generate responses that deviate from the intended context of the problem. Moreover, the task explanations often lacked the essential information needed to complete them successfully. These shortcomings arose because the dataset was compiled by numerous non-experts through *Amazon Mechanical Turk*.

In addition to the issues highlighted in Figure 2, there were further cases of inconsistency throughout the dataset. These inconsistencies involved not only color but also shape representations and grid manipulation operations. The presence of these multiple issues complicates the process of generating new datasets based on LARC, emphasizing the challenges of relying on flawed data sources.

To address these issues, we conducted a refining process to enhance quality. This process prioritized ensuring consistency in expressions and rectifying erroneous representations. Figure 2 provides an overview of this refining process.

Designing Distractors with ChatGPT4 Based on the given output description of refined LARC, we generated four distractors through ChatGPT4, as illustrated in Figure 3. However, allowing unrestricted generation of distractors led to issues

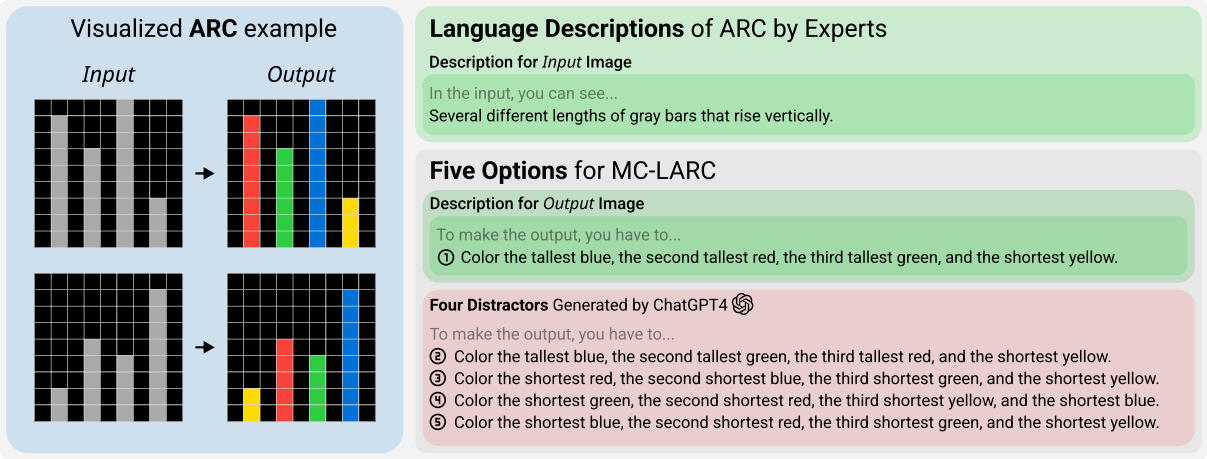

Figure 3: MC-LARC structure. Visualized ARC example (blue) with five options - one correct solution and four distractors. Refined LARC by experts (green) and GPT-4 generated distractors based on refined LARC description (red). The solver must infer common rules from the ARC example to select the best-matching option.

such as creating out-of-context choices unrelated to the task. To address this problem, we improved by adding constraints during the prompt level. The constraints added to the prompt are as follows:

- **In Context Vocabulary**: To generate plausible distractors, it was necessary to limit the expressions within the context that aligns with the ARC domain. To achieve this, two contextual constraints were imposed. One involved adding descriptions about the ARC environment, while the other entailed mentioning specific words that should not be used.
- **Length of Options**: When generating distractors for lengthy options, there were cases where LLM produced relatively short options, leading to easily solvable problems. Therefore, we restricted the LLM to generate incorrect options of similar lengths to the correct options.
- **Format**: When creating distractors, we ensured that the opening phrases of the sentences exactly matched the correct answer option's *'To make the output, you have to...'*. If the opening phrases of the incorrect options vary, it could lead to selecting the correct answer based on the format rather than the meaning of the sentence.

We analyzed to determine the extent of MC-LARC's refinement process impact. Table 1 illustrates the word count differences between correct and incorrect options before and after refinement. Notably, adding constraints significantly decreased the disparity in average word count and variance between correct and incorrect options. This reduction in disparity serves to mitigate potential shortcuts based on option length.

Table 1: Word count statistics before and after refinement, comparing correct and incorrect options. The mean word count for incorrect options increased from approximately 29 to 37, greatly reducing the gap with the correct options' 39 and making them more similar.

Word Count Statistics	Before	After
Correct	39.73 ± 28.13	$\mathbf{39.08 \pm 26.61}$
Incorrect	29.01 ± 18.93	$\mathbf{37.34 \pm 22.40}$

Table 2: Similarity metrics before and after refinement, comparing correct and incorrect options. The increase in Jaccard similarity (Leskovec et al., 2020) and the decrease in Levenshtein distance (Levenshtein, 1966) indicate that the similarity between options has improved after the refinement process.

Similarity Metrics			
Metric	**Statistic**	**Before**	**After**
Jaccard Similarity	Mean	0.404	**0.777** ↑
	Variance	0.021	0.017
Levenshtein Distance	Mean	0.439	**0.129** ↓
	Variance	0.021	0.009

Table 2 presents the Jaccard similarity and Levenshtein distance between correct and incorrect options. A higher Jaccard similarity indicates greater textual similarity, while a lower Levenshtein distance signifies increased similarity. The increased

similarity between correct and incorrect options post-refinement suggests a convergence in sentence structure. This convergence helps prevent shortcuts based solely on option formatting.

4 Experiments

4.1 Evaluating MC-LARC's Efficacy

This section evaluates the effectiveness of MC-LARC in assessing lower-level cognitive skills, focusing on understanding and application in reasoning tasks. For this purpose, we presented all MC-LARC problems to the ChatGPT4V model five times, resulting in an accuracy of about 76% for the total 400 tasks. Considering that the accuracy of LLMs on ARC tasks is around 10% (Qiu et al., 2024), this is certainly a high score. To investigate the implications of this difference, we conducted an additional experiment where we asked the LLM for the reasoning behind selecting each option. Results revealed a strong correlation between answer accuracy and reasoning validity, as illustrated in Figure 4. Instances of correct answers with incorrect explanations, or incorrect answers with valid reasoning, were negligible. This indicates reduced errors such as generating correct answers through incorrect reasoning processes or providing inconsistent answers, which often occur when LLMs directly solve ARC tasks (Lee et al., 2024). Furthermore, the LLM's explanations reflect its understanding of the ARC task by following the problem-solving process within the options. While the multiple-choice format of MC-LARC does not directly assess the complex pattern recognition, rule extraction, abstraction, and application to new situations required by ARC, it allows for the observation of intermediate steps in the ARC problem-solving process. Consequently, MC-LARC allows for a closer analysis of the LLM's fundamental understanding and reasoning process.

To validate MC-LARC's reliability as an evaluation metric, we assessed consistency in LLM responses. Table 3 presents these results. KR-20 and Cronbach's Alpha values exceeding 0.9 indicate strong internal consistency. The ANOVA analysis, showing a small F-statistic and p-value above 0.05, demonstrates consistent evaluation across items without significant variation. These findings confirm MC-LARC as a highly reliable test for assessing respondents' abilities.

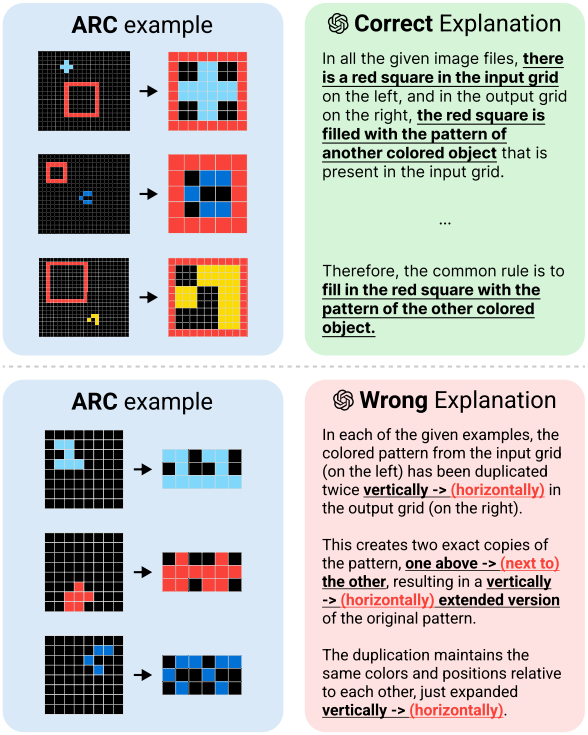

Figure 4: A result of requesting an explanation of the experiments with provided images. (Upper) Shows an example where the answer to MC-LARC is correctly chosen. (Lower) Demonstrates the incorrect answers due to failure to infer the correct solution.

Table 3: Analysis of response consistency reliability in experiments with and without ARC images. Based on the LLM solving 400 MC-LARC tasks five times. Higher KR-20 (Kuder and Richardson, 1937) and Cronbach's Alpha (Cronbach, 1951) indicate greater internal consistency. Higher ANOVA p-values and lower F-statistics (Scheffe, 1999) suggest less significant differences between attempts, indicating more consistent responses across trials. ↑ and ↓ arrows indicate better consistency in the respective condition.

Metric	With Image	Without Image
KR-20	0.918	0.922 ↑
Cronbach's Alpha	0.917	0.921 ↑
ANOVA p-value	0.862 ↑	0.712
ANOVA F-statistic	0.324 ↓	0.532

4.2 Problems on Augmentation

However, we discovered an interesting finding: LLMs use a shortcut to solve MC-LARC. As shown in Figure 6, we uncovered this fact through a comparative experiment analyzing the results and processes of LLMs solving the problems with and without providing the ARC images. MC-LARC should be solved by inferring the rule from the given images and choosing the correct option, but

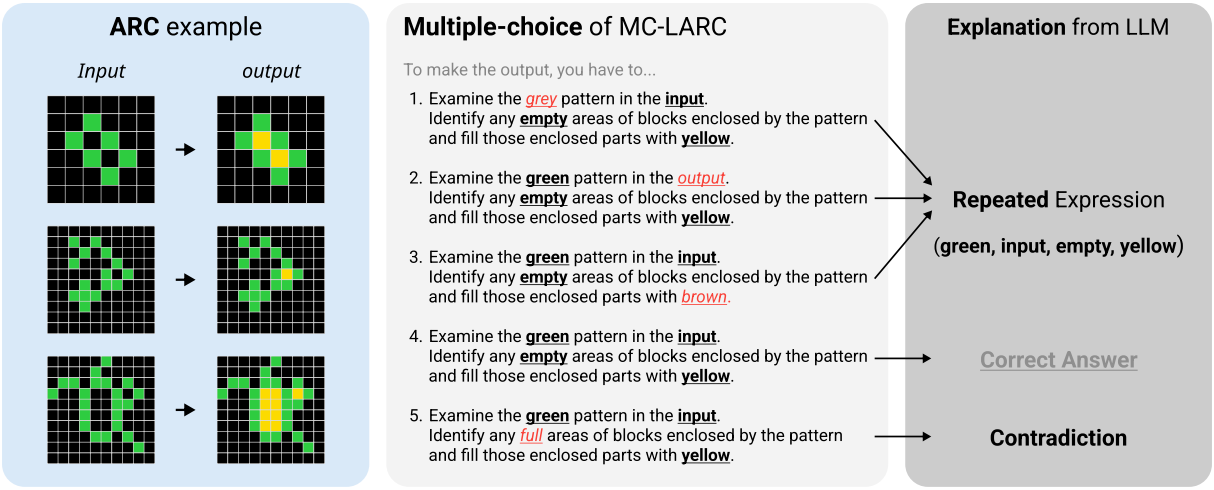

Figure 5: LLM problem-solving methodology in MC-LARC. 1) Option analysis: Identifies recurrent expressions. 2) Vocabulary consistency check: Excludes options with divergent terminology 3) ARC domain compatibility assessment: Eliminates options with semantic contradictions incompatible with the ARC task.

Table 4: Comparison of MC-LARC solving performance between ChatGPT4V and humans, with and without images. It shows mean accuracy from five experiments. For more detailed information on the human evaluation, please refer to Section 5.2.

Image	Solver	Accuracy (%)
With	ChatGPT4V	76.05 ± 1.34
	Human	90.75 ± 2.85
Without	ChatGPT4V	**64.61 ± 2.17**
	Expected Value	**20.00**

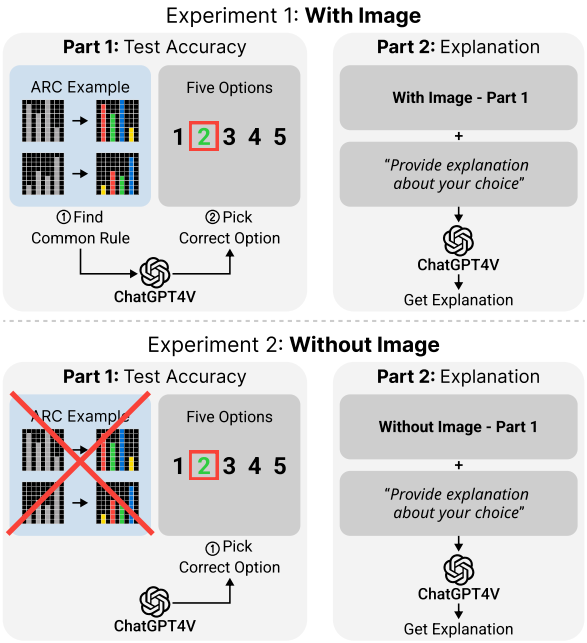

Figure 6: Experimental design overview. (Upper) Image-based experiment: Utilizes visualized ARC examples. (Lower) Text-only experiment: Excludes visual aids. Both experiments comprise two parts. Part 1: ChatGPT4 MC-LARC problem-solving for accuracy assessment. Part 2: Solution explanation alongside problem-solving, building on Part 1 tasks.

the LLM achieved an accuracy of 65% even when the task was provided without images.

To analyze how the LLM solved MC-LARC without the problem images, we additionally asked the LLM to explain the reasoning behind its answers. As shown in Figure 5, we found that the LLM inferred the correct option by 1) choosing the option with the most repeated expressions and 2) eliminating self-contradictory options.

We point out two problems in the generation process: First, we notice an unintended pattern when LLM generates the four distractors from the correct answer. The correct option often contained words that appeared most frequently across all choices. As shown in Figure 5, the distractors describe terms that differ from the common keywords shared by the other four options, making it easier to identify them as incorrect. This linguistic pattern could unintentionally hint at the correct answer. Second, not providing image and context information for option generation led to contradictory expressions, and we confirmed that the LLM identified these distractors by detecting semantic contradictions by comparing the options. Therefore, from this experiment, we can conclude that to evaluate reasoning ability fairly, the process of generating choices should be improved to avoid providing additional information that could serve as a shortcut.

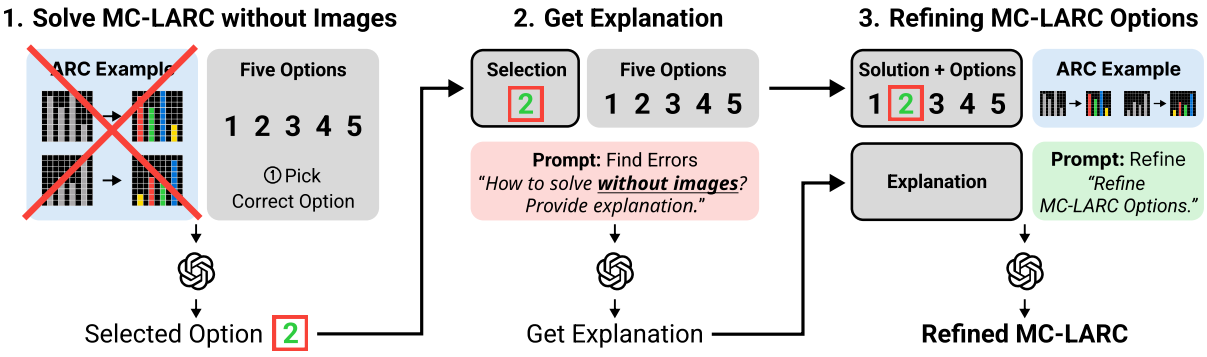

Figure 7: The self-feedback process for improving MC-LARC quality. 1) Initial problem-solving: LLM solves MC-LARC without visual input. 2) Solution justification: LLM explains the reasoning process and finds errors of options for Stage 1. 3) Comprehensive revision: LLM refines MC-LARC options considering ARC example images, correct answers (solution), options, and Stage 2 explanation.

4.3 Improving Quality: Self-Feedback Framework

From the two experiments above, we confirmed that converting to a multiple-choice format has advantages as an inference problem in two aspects: 1) providing additional information to solve the reasoning problem, and 2) allowing for a more transparent evaluation of the reasoning process. However, we also found cases where unintended shortcuts were discovered, and to address this issue, the process of augmenting choices needs to be improved.

We conducted an additional experiment comparing the original MC-LARC with an improved version using a self-feedback process inspired by a previous study (Wang et al., 2024). As illustrated in Figure 7, the self-feedback process consists of three stages. First, the problem is solved without the image. Next, the problem-solving process is explained without the image to identify potential shortcuts. Finally, new options are generated that address the identified shortcuts. This framework enhances the quality of options without adding explicit constraints.

Table 5: LLM performance on MC-LARC: Comparing image presence and refinement effects on accuracy and shortcut reduction. Refinement reduces 'without image' accuracy towards ideal 20%, indicating fewer shortcuts.

Image	Version	LLM Accuracy (%)
With	Before	76.05 ± 1.34
	After	62.50 ± 2.32
Without	Before	64.61 ± 2.17
	After	$\mathbf{43.75 \pm 1.55} \downarrow$

As shown in Table 5, the significant decrease in accuracy without image after refinement (from 64.61% to 43.75%) suggests a substantial reduction in shortcuts. However, after the revision, the average accuracy when an image was provided dropped from 76.05% to 62.5%. This seems to be due to the increased difficulty of the options, as their similarity increased while reducing shortcuts. In summary, after applying the self-feedback framework, the accuracy gap between image-present and image-absent conditions widened, indicating improved option quality and reduced reliance on contextual cues. Conversely, the decrease in accuracy for image-present conditions post-revision suggests increased ambiguity among options.

5 Discussion

5.1 Criteria on Good Option and Bad Option

In essence, the central challenge revolves around distinguishing between what constitutes a good problem and what does not. Before we can enhance the process of generating answer choices, we must first address this fundamental question: What are the distinguishing factors between high-quality and low-quality answer options?

As we examined the augmented choice examples generated by the LLM, we could categorize the choices into three levels of quality, as shown in Figure 8. The best choices modified the core part of the problem that fits the context. In ARC, the core is the part where a change occurs between images, so in the given examples, completing a square by filling in orange pixels is the core. Thus, choices that question the change to orange can be considered the best type of choice. Next, choices that were possible to predict from the input image

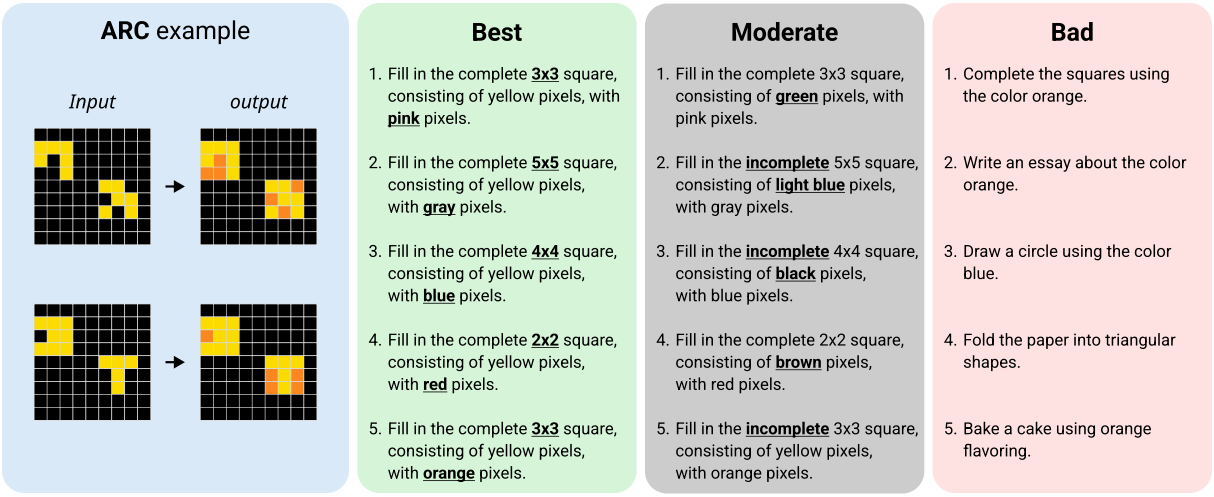

Figure 8: Three examples of multi-choice options augmented differently by the LLM. The given problem is to fill in an object with holes with the color orange to make a 3×3 square, where the size of the square and the color are the core aspects of the problem. The good example demonstrates an understanding of the core of the problem and provides consistent variations, while the poorer examples increasingly include choices that are unrelated to the problem and inconsistent.

but did not capture the core of the problem were of moderate quality. Examples include using colors not present in the input image or specifying grid sizes that were not present. Finally, choices that included cases that cannot occur in the ARC domain at all were the worst. Commands like 'Write an essay' are irrelevant to ARC and do not require any reasoning process to solve the problem, making them poor choices.

Therefore, good text descriptions should 1) include the core of the problem in the choices, and 2) be consistent within the context of the problem. Identifying the criteria in form and content needed to generate good choices during the augmentation process is the contribution of this study.

5.2 Human Evaluation of MC-LARC

To assess the efficacy of MC-LARC in capturing human-level reasoning, we conducted a comprehensive evaluation involving human participants. We recruited 8 undergraduate interns from our laboratory to evaluate the initial MC-LARC version. To manage cognitive load, we divided the 400 MC-LARC tasks into 8 sets of 50 tasks, assigning one set to each participant. Table 6 shows the results.

The human evaluation yielded key insights: aggregating results from all participants, we estimated a high overall accuracy of 90.75%, with individual performances ranging from 72% to 100%. This approach enabled assessment of the full dataset while managing participants' cognitive load.

Table 6: Human performance on MC-LARC: Individual accuracy on 50-task subsets and overall result

ID	Solved Tasks	Accuracy (%)	
1	1–50	94.00 ± 6.82	[87.18–100.00]
2	51–100	72.00 ± 12.89	[59.11–84.89]
3	101–150	86.00 ± 9.56	[76.04–95.96]
4	151–200	86.00 ± 9.56	[76.04–95.96]
5	201–250	100.00 ± 0.00	[100.00–100.00]
6	251–300	96.00 ± 5.63	[90.37–100.00]
7	301–350	94.00 ± 6.82	[87.18–100.00]
8	351–400	98.00 ± 4.02	[93.98–100.00]
	Overall	**90.75 ± 2.85**	**[87.90–93.60]**

To gain insights into how human performance varies with task complexity, we surveyed the participants on the difficulty of each MC-LARC task and analyzed the accuracy across different difficulty levels. Table 7 presents these results. Performance generally declined with increasing difficulty, particularly at the highest level.

Notably, human participants outperformed the LLM (ChatGPT-4V), underscoring MC-LARC's effectiveness in capturing human-level reasoning and its potential as a challenging benchmark for AI systems. These findings highlight MC-LARC's value in evaluating and advancing AI capabilities, with future work aimed at analyzing LLM performance across different difficulty levels.

Table 7: Human performance across MC-LARC difficulty levels: Accuracy decreases with complexity

Difficulty	Total	Correct	Accuracy (%)
1	155	149	96.13
2	102	91	89.22
3	72	67	93.06
4	41	38	92.68
5	30	18	60.00
Overall	**400**	**363**	**90.75**

5.3 Comparative Analysis of Language Description Quality

To evaluate the utility of MC-LARC, we designed an experiment to assess its effectiveness in solving ARC tasks and as a source of labeled data for learning. The language descriptions in MC-LARC options, including correct and incorrect analogies, can serve as valuable resources. We examined how effectively MC-LARC aided in directly solving ARC problems.

The experiment was structured to evaluate the LLM's program synthesis capabilities on ARC tasks. We provided the input and output of ARC tasks along with a set of Python functions capable of solving each problem. The LLM's objective was to identify the correct combination of functions to solve the given problem. In addition to the solution process from MC-LARC, we incorporated data from Fast and Flexible (Johnson et al., 2021) and LARC datasets (Acquaviva et al., 2022), supplying step-by-step functions written in Python, explanations of ARC problems, and the corresponding task's input and output. The experiment was conducted 10 times on 20 problems common to all three datasets, enabling a comprehensive comparison of the LLM's performance across different approaches and allowing us to assess the relative value of MC-LARC's language descriptions as learning labels. Please refer to Section 3.2 and B.2 of the Appendix in (Lee et al., 2024) for detailed information on the experimental setting.

Table 8 presents the compositionality performance of LLMs across different benchmarks. The results indicate that the highest average accuracy was observed when the LLM provided MC-LARC descriptions. This suggests that the refined correct options of MC-LARC more effectively capture the key aspects of problem-solving in ARC tasks.

MC-LARC descriptions improved LLM performance compared to other datasets and baselines, highlighting the value of well-crafted language descriptions in enhancing compositionality. This implies that the ability to generate textual information not explicitly provided during inference is crucial, as it helps the LLM infer missing context and approach with a deeper understanding. Also, MC-LARC's inclusion of incorrect options enables contrastive learning. These findings emphasize the importance of diverse, high-quality language descriptions in improving LLM understanding and problem-solving, particularly for compositional reasoning tasks, positioning MC-LARC as a valuable resource for advancing AI learning techniques.

Table 8: Compositionality performance for ARC problems. When MC-LARC was provided, it was observed that the highest accuracy rates were achieved.

Metric	Accuracy (%)
No Description	8.0 ± 0.09
Fast and Flexible (Johnson et al., 2021)	8.0 ± 0.09
LARC (Acquaviva et al., 2022)	13.0 ± 0.11
MC-LARC	$\mathbf{14.5 \pm 0.12}$

6 Conclusion

To overcome the limitations of the existing ARC in measuring inferential reasoning ability, we created a new multiple-choice dataset called MC-LARC. As a result, the multiple-choice format allowed for a clearer analysis of logical flow during problem-solving and provided support for the solver's reasoning abilities. However, in an additional control experiment without images, we found that the LLM solved problems by finding shortcuts instead of using reasoning abilities. This highlights the regulation needed when using LLMs to synthesize multiple-choice questions. Based on these findings, we introduce a self-feedback framework to address shortcuts. This framework represents our distinctive approach, using LLMs to generate proper descriptions, thereby mitigating the shortcut problem.

These findings have several important implications. Firstly, they offer valuable insights into the appropriate methods for evaluating inferential reasoning, demonstrating the potential of using multiple-choice questions for this purpose. Secondly, by identifying the constraints to consider when using LLMs to synthesize multiple-choice questions, this research proposes a framework for the development of more sophisticated and automated high-quality description generators.

7 Limitation

Our study has two main limitations. First, despite improvements, shortcuts may still persist. Second, there is a lack of metrics to measure the quality of the options. We observed that even after enhancement through self-feedback, approximately 40% of the problems could be solved without images. However, these issues are inherent limitations of multiple-choice questions (Alagumalai and Curtis, 2005), and therefore, do not undermine the fundamental purpose of MC-LARC to assess cognitive features of LLMs such as understanding and application, which are difficult to confirm solely through solving ARC problems.

Secondly, our current analysis is limited to the accuracy of LLMs. In existing test theory, metrics such as discrimination are used to evaluate the quality of options. This requires the use of various LLMs and analysis of human cases. Nonetheless, this study lays the foundation for identifying cognitive features that cannot be confirmed through ARC alone, with significant potential for future expansion.

Acknowledgements

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 were supported by the Accelerate Foundation Models Research Initiative, Microsoft.

References

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*.

Sivakumar Alagumalai and David D. Curtis. 2005. *Classical Test Theory*. Springer.

Lorin W. Anderson, David R. Krathwohl, Peter W. Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul R. Pintrich, James Raths, and Merlin C. Wittrock. 2001. *A Revision of Bloom's Taxonomy of Educational Objectives*. Pearson.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*, pages 2425–2433.

Patricia Armstrong. 2010. Bloom's Taxonomy. *Vanderbilt University Center for Teaching*, pages 1–3.

Mikhail Moiseevich Bongard. 1968. *The Recognition Problem*. Foreign Technology Div Wright-Patterson AFB Ohio.

Susan M. Case and David B. Swanson. 1998. *Constructing Written Test Questions for the Basic and Clinical Sciences*. National Board of Medical Examiners Philadelphia.

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

Lee J. Cronbach. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16(3):297–334.

Lin Ding and Robert Beichner. 2009. Approaches to Data Analysis of Multiple-Choice Questions. *Physical Review Special Topics-Physics Education Research*, 5(2):020103.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *CVPR*, pages 2758–2766.

Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. 2024. MARVEL: Multidimensional Abstraction and Reasoning through Visual Evaluation and Learning. In *NeurIPS*.

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*.

Alark Joshi, Chandana Srinivas, Elif E. Firat, and Robert S. Laramee. 2024. Evaluating the Recommendations of LLMs to Teach a Visualization Technique Using Bloom's Taxonomy. *Electronic Imaging*, pages 1–8.

Subin Kim, Prin Phunyaphibarn, Donghyun Ahn, and Sundong Kim. 2022. Playgrounds for Abstraction and Reasoning. In *NeurIPS Workshop on nCSI*.

G Frederic Kuder and Marion W Richardson. 1937. The Theory of the Estimation of Test Reliability. *Psychometrika*, 2(3):151–160.

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

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. *Mining of Massive Data Sets*. Cambridge University Press.

V Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Proceedings of the Soviet Physics Doklady*.

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*.

Edward Palmer, Peter Devitt, et al. 2006. Constructing Multiple Choice Questions as a Method for Learning. *Annals-Academy of Medicine Singapore*, 35(9):604.

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*.

Henry Scheffe. 1999. *The Analysis of Variance*, volume 72. John Wiley & Sons.

Hassan Shojaee-Mend, Reza Mohebbati, Mostafa Amiri, and Alireza Atarodi. 2024. Evaluating the Strengths and Weaknesses of Large Language Models in Answering Neurophysiology Questions. *Scientific Reports*, 14(1):10785.

Apostolos Spanos et al. 2024. BloomGPT: Using ChatGPT as Learning Assistant in Relation to Bloom's Taxonomy of Educational Objectives. In *Conference Proceedings. The Future of Education 2024*.

Rashmi Vyas and Avinash Supe. 2008. Multiple Choice Questions: a Literature Review on the Optimal Number of Options. *Natl Med J India*, 21(3):130–3.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-Agents Enhances Large Language Model Capabilities. *arXiv:2406.04692*.

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*.

Appendix

A Additional Experiments on MC-LARC Option Reliability

To further investigate the statistical variations in MC-LARC option reliability and accuracy under different constraint conditions, we conducted additional experiments. Table 9 summarizes our findings.

[1]Constraint key: g - grid system explanation; c - color description; s - vocabulary restriction; f - figure provided; o - caution against repeated expressions; t - caution against contradictions

Table 9: Reliability metrics and ANOVA results by constraint condition[1]

Constraint	KR-20	Cronbach's α	F-statistic
cs	0.890	0.872	0.535
gc	0.842	0.818	0.622
gs	0.828	0.803	0.456
gcs	0.870	0.850	0.414
gcsf	0.915	0.900	1.264
gcsotf	0.673	0.631	0.675

Our analysis revealed high reliability for problems with provided constraints. However, no statistically significant differences were observed across the various constraint conditions. This lack of substantial variation may be attributed to the absence of appropriate quantitative indicators reflecting option quality.

B Analysis of Potential Shortcuts in MC-LARC Problem-Solving

To investigate whether LLMs exploit shortcuts based on formal aspects of the tasks, such as word count or ARC image colors, we conducted t-tests comparing tasks groups categorized by accuracy rates into well-solved (easy) and poorly-solved (difficult) groups. If shortcuts existed in formal aspects, we would expect to observe significant statistical differences in specific attributes (e.g., number of words, number of pixels) between these groups.

B.1 Methodology

We divided the 400 MC-LARC tasks into two groups based on the LLM's performance:

- Well-solved (Easy) tasks: 291 tasks were solved correctly 4 times or more out of 5 trials
- Poorly-solved (Difficult) tasks: 109 tasks were solved correctly 3 times or less out of 5 trials

We then conducted t-tests to compare various metrics between these groups, both when images were provided and when they were not.

B.2 Results

Table 10 present the results of our t-tests for various metrics. The analysis revealed no statistically significant differences between easy and difficult tasks across all measured metrics. This held both when images were provided and when they were not.

Table 10: Comparison of task attributes between easy and difficult MC-LARC tasks for LLM (with image information): Analysis of input/output features and option word counts shows no significant differences.

Metric	**Easy**	**Difficult**	t**-statistic**	p**-value**
Number of Input	3.21	3.38	-1.603	0.110
Number of Output	3.21	3.38	-1.603	0.110
Average Input Pixel Numbers	148.21	132.60	0.911	0.363
Average Input Color Types	3.55	3.46	0.441	0.660
Average Output Pixel Numbers	118.37	92.76	1.853	0.065
Average Output Color Types	3.39	3.28	0.685	0.494
Correct Option Word Count	40.17	36.16	1.527	0.128
1st Incorrect Option Word Count	38.38	35.54	1.184	0.238
2nd Incorrect Option Word Count	37.86	35.18	1.117	0.265
3rd Incorrect Option Word Count	37.93	35.28	1.099	0.273
4th Incorrect Option Word Count	38.10	35.51	1.071	0.285

These results suggest that the number of words and other formal aspects of the problems do not currently function as shortcuts when LLM solves MC-LARC. This finding has important implications:

1. It indicates that shortcuts cannot be resolved by simply controlling formal aspects such as word count or format in a mechanical way.

2. It highlights the complexity of addressing shortcuts in language model performance, suggesting that more sophisticated approaches may be necessary.

3. The lack of significant differences in formal aspects between easy and difficult tasks implies that the LLM's performance is likely based on more nuanced features of the problem descriptions or underlying reasoning processes.

These insights contribute to our understanding of LLM behavior in complex reasoning tasks and underscore the challenges in identifying and mitigating shortcut learning in such contexts.

C Potential Enhancements to Multi-Choice Generation Methodology

While the experimental results confirmed that the multiple-choice problem format provided sufficient additional information to adequately assess *Understand* and *Apply* aspects, the issue of finding shortcuts during the solving process was raised. This problem is not unique to LLM evaluation. The issue of imbalance among options in multiple-choice questions has already been raised in classical test theory (Alagumalai and Curtis, 2005). The following are suggestions for improving the options in MC-LARC:

- **Option Quality Improvement**: The multiple-choice evaluation method has been criticized for the existence of shortcuts such as *Logical cues*, *Long correct answer*, *Word repeats*, and *Convergence strategy*, even in the case of humans (Case and Swanson, 1998). It has also been pointed out that when there is a lack of discrimination power, the quality of the options decreases. The most intuitive way to address this issue is for humans to consider constraints when creating options.

- **Modification on the Benchmark Format**: Not only the content of the options but also the format of the options can affect the benchmark. Currently, MC-LARC follows a format where one correct answer option is chosen among five options. On the other hand, another study reported that the selection ratio between options remained similar when there were four or three options compared to five options (Vyas and Supe, 2008). It is also noteworthy that problems with multiple correct answers tend to be more difficult than those with a single correct answer (Case and Swanson, 1998). However, it is not yet known how these various multiple-choice formats differ for LLMs, and therefore, they need to be considered as hyperparameters in the future.

- **Changing the Evaluation Objective**: Modifying the content of the multiple-choice options to measure various areas of reasoning

such as application and creation is another possible improvement. Currently, the options in MC-LARC are focused on finding the correct way to solve the ARC task, which is aimed at assessing the understanding of the task. To extend the assessment to other reasoning abilities, the application and creation stages of the task need to be evaluated. Converting the problem into a multiple-choice format where images are selected instead of answer texts, similar to MARVEL (Jiang et al., 2024), could be one possible way to shift the problem format to the creation stage. To transition to the application stage, instead of using an entire problem description, it may be necessary to consider separating the steps required to solve the problem and have the option to select steps that are not necessary for solving the given ARC task.

D Potential Enhancements in the Evaluation Methodology

One of the current limitations of MC-LARC is the lack of sufficient evaluation metrics for the proposed benchmark. Therefore, it is difficult to assess how much the addition of multiple-choice has contributed to securing intermediate reasoning stages leading up to ARC, and how well the options are constructed. The following describes existing methods for evaluating options:

- **Using Scoring Models**: Ding and Beichner (2009) has proposed statistical and numerical methods for evaluating the quality of multiple-choice questions (MCQs). They propose three methods for individual item evaluation (*Item Difficulty Level*, *Item Discrimination Index*, *Point Biserial Coefficient*) and two methods for overall test evaluation (*Kuder-Richardson Reliability Index*, *Ferguson's Delta*). *Item Difficulty Level* and *Item Discrimination Index* measure item difficulty and discriminative power, while *Point Biserial Coefficient* assesses each item's appropriateness by comparing item scores with the total test score. The *Kuder-Richardson Reliability Index* determines whether the test is suitable for individual or group assessments, and *Ferguson's Delta* measures the test's ability to distinguish between varying levels of proficiency. Additionally, they introduce clustering analysis for analyzing respondent patterns and model usage. Therefore, using metrics to measure the quality of MCQs is one method for improving MC-LARC.

- **Comparison with Human-Created Questions**: One issue with the current MC-LARC is that both question generation and evaluation are done through a single model, ChatGPT4V. This evaluation approach does not reveal whether MC-LARC can be properly evaluated on other models, including other LLMs. In existing test theory, to compare with human-created options, a large number of people directly participated in the evaluation to minimize errors as much as possible (Palmer et al., 2006). Similarly, 1) three or more people can evaluate whether there are errors in the options, and 2) the quality of the options can be compared with human-created questions.