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

Anonymous ACL submission

Abstract

As the reasoning abilities of artificial intelligence gain more attention, generating reliable benchmarks to evaluate reasoning capabilities is becoming increasingly important. The Abstract and Reasoning Corpus (ARC) is one of the introduced reasoning benchmarks, providing challenging problems that artificial intelligence has yet to solve. While ARC has been recognized for assessing reasoning abilities, it has a limitation in that its evaluation method through generation fails to consider other aspects of assessment. Bloom's taxonomy, widely known in education, argues that good evaluation methods should evaluate the six stages of *Remember*, *Understand*, *Apply*, *Analyze*, *Evaluate*, and *Create* in a step-by-step manner. To make ARC, which primarily evaluates the *Create* stage, suitable for assessing stages like *Understand* and *Apply*, we developed MC-LARC. This new multiple-choice format fits well on evaluating large language models (LLMs) across different cognitive stages. We evaluated the analogical reasoning abilities of ChatGPT4V with MC-LARC, confirming that 1) a multiple-choice format can support the language model's reasoning capabilities and 2) facilitate evidence analysis. However, we noticed LLMs relying on shortcuts when tackling MC-LARC. By analyzing this, we identified areas to consider in multiple-choice synthesis and specified criteria for what constitutes good choices based on these findings.

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 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 75%. This indicates that the options in MC-LARC have served a supportive role in the inference of LLMs, which are more aligned with language generation and comprehension than image processing. 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. Based on this analysis, it was confirmed that when synthesizing data into a multiple-choice format using LLMs, sufficient and accurate context information should be provided to avoid unnecessary additional information. Additionally, this analysis established criteria for what constitutes good multiple-choice options.

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*).

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.

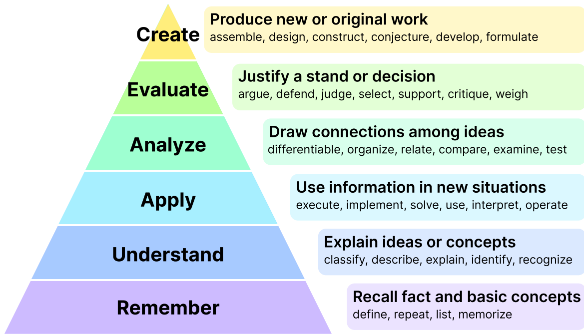

Figure 1: The six cognitive skills in Bloom's Taxonomy. These skills begin with basic tasks like recalling facts and remembering concepts at the pyramid's base, progressing to creating original work based on a comprehensive understanding of a specific concept at the top. Image credits: Center for Teaching,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

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 Benchmark for Analogy 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 LARC 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, 1967), 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 are 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 Methodology

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.

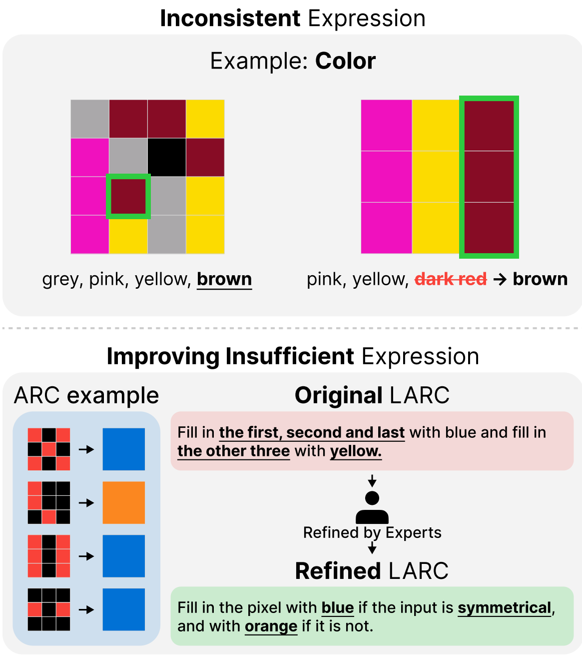

Figure 2: Two main issues of LARC. (Upper part) There are instances where different expressions are used for the same concept within LARC. For example, some LARC expressions describe brown as "dark red". (Lower part) This task involves identifying the symmetry of the input grid to predict the output image result. However, some original LARC expressions provide insufficient information necessary for ARC problem-solving. These have been revised to contain sufficient and accurate information by experts.

Refining process The original LARC exhibited significant quality issues, as evidenced by Figure 2. These issues appeared primarily in 1) inconsistencies across expressions for the same concept and 2) a lack of information in the provided explanations. For instance, the upper part of Figure 2 illustrates different representations for the same concepts, leading to confusion. Additionally, the explanations accompanying the tasks often omitted crucial information necessary for their successful completion. These issues emerged as a consequence of the dataset's compilation by numerous non-experts using 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

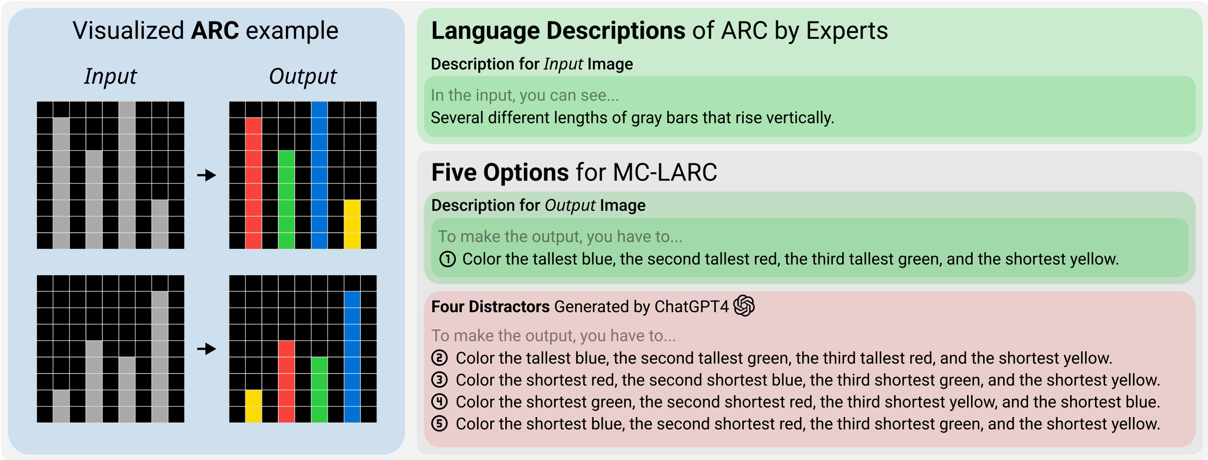

Figure 3: The composition of MC-LARC. It consists of a visualized ARC example and five multiple-choice options. The five multiple-choice options consist of the correct solution and four distractors. (Blue part) It visualizes ARC represented as a 2D matrix. (Green part) It is LARC refined manually by experts. (Red part) Using ChatGPT4, four distractors were generated from the output description (Red boundary) of the refined LARC. To solve MC-LARC, the solver must identify common rules from the visualized ARC example and select the one option from the *Five Options for MC-LARC* that best describes those rules.

process to enhance quality. This process prioritized ensuring consistency in expressions and rectifying erroneous representations. Figure 2 provides an overview of this refining process.

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

As shown in Figure 7, before constraints were added, the model generated options that were either completely irrelevant to the ARC problem context or altered parts that were not core concepts. These were classified as either `bad` or `moderate`. However, after the constraints were applied, the model did not produce any `bad` options, and the options were classified only as `best` or `moderate`. Despite this improvement, the model still faces the challenge of not being able to produce `best` options for all tasks.

4 Experiments

To verify that the augmented multiple-choice options generated by the LLM did not inadvertently reveal more information than intended, we conducted a control test, as illustrated in Figure 4, where the LLM was presented with only the options, devoid of any accompanying images. If the options were crafted appropriately and free from informational bias, the LLM's expected accuracy rate would approximate 20%. Additionally, this image-free experiment required the LLM to justify its choice for each option.

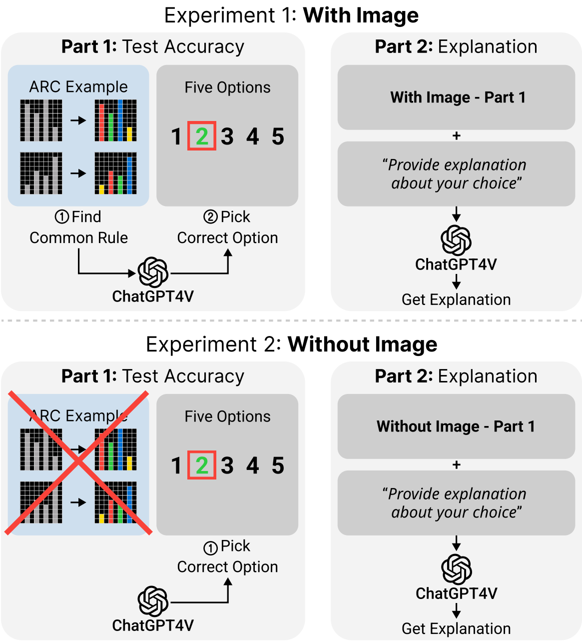

Figure 4: Overview of the conducted experiments. The upper part illustrates the first experiment, which includes visualized ARC example images, while the lower part depicts the second experiment, which does not include these images. Each experiment is divided into two parts. In Part 1, ChatGPT4 is tasked with solving the MC-LARC to measure accuracy. In Part 2, it is requested to provide explanations for its choices, in addition to completing the tasks from Part 1.

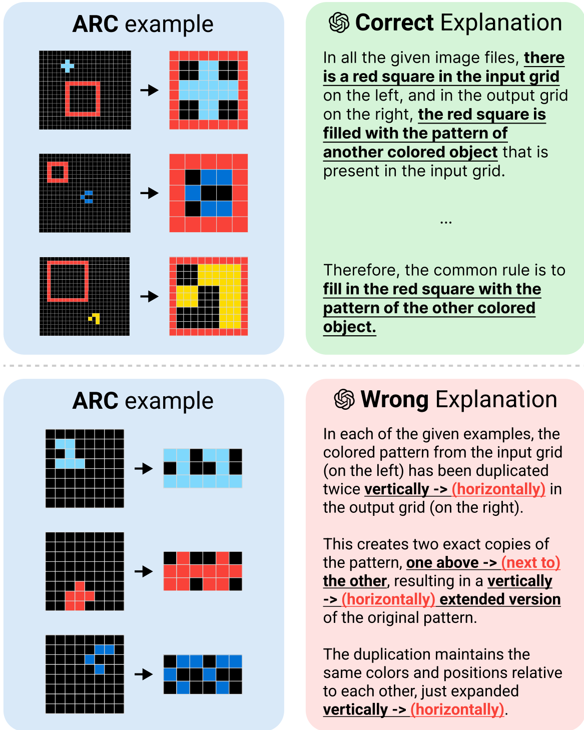

Figure 5: A result of requesting an explanation of the experiments with provided images. (Upper part) It shows an example where the answer to MC-LARC is correctly chosen. (Lower part) It demonstrates the incorrect answers due to failure to infer the correct solution.

4.1 Influence of Multiple Choices

Table 1: A table summarizing the results of experiments where ChatGPT4V solved MC-LARC five times. It shows statistics on the accuracy and Krippendorff's Alpha score. The statistics show the mean, standard deviation, and 95% confidence interval for the accuracy. Krippendorff's Alpha score evaluates whether ChatGPT4V's responses are reliable across the five repeated experiments.

Category	Mean (%)	Std.	95% CI (%)	Alpha
With images	75.81	1.11	74.93 - 76.70	0.8329
Without images	64.61	1.75	63.08 - 66.14	0.7995

For the MC-LARC, we asked the ChatGPT4V model 5 times per problem, and as shown in Table 1, the accuracy of correctly answered tasks out of the total 400 tasks was about 75%. Considering that the accuracy of LLMs on ARC tasks is around 10% (Qiu et al., 2024), this is certainly a high score. Additionally, Krippendorff's Alpha score of approximately 0.83 confirmed that the LLM was consistently reasoning the answers.

To further evaluate the reasoning process of the LLM, we additionally asked for the reason behind selecting each option. As a result, there were cases where both the answer and the reasoning process were correct or both were incorrect, but there were almost no cases where the answer was correct but the explanation was wrong, or where the answer was wrong but the explanation was correct. This indicates a decrease in the errors of generating correct answers through incorrect reasoning processes or giving inconsistent answers, which tend to occur when LLMs directly solve ARC tasks (Lee et al., 2024). Therefore, even when multiple-choice options, including incorrect options along with the answer description, were provided, we could confirm that the LLM's reasoning ability was partially improved.

4.2 Problems on Augmentation

However, there were indications that the LLM found a shortcut when solving MC-LARC. MC-LARC should be solved by inferring the rule from the given images and choosing the correct option, but the LLM achieved an accuracy of 65% even when the task was provided without images. The Krippendorff's Alpha score was also 0.79, not

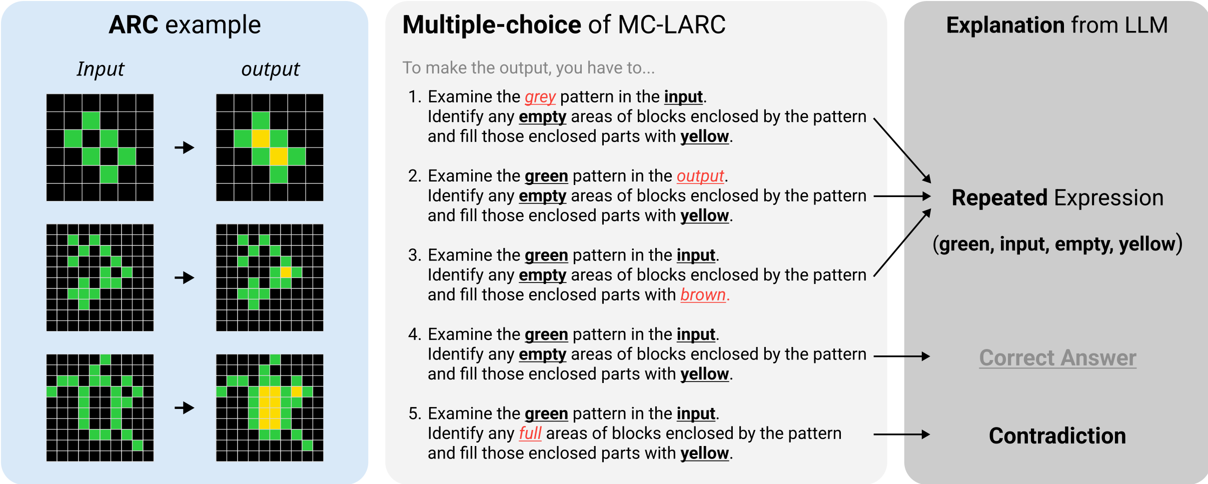

Figure 6: Example of an experiment without an image. When given five options, the LLM solves the problem by analyzing them in the following manner. By examining the options, the LLM identifies repeated expressions and excludes the options that use different vocabulary from the others. Additionally, it excludes options that cannot be represented in the ARC grid by identifying semantic contradiction within the sentences themselves.

much lower than the experiment with images provided. This can be understood as evidence that the LLM found a consistent logic for getting the correct answers.

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 6, we found that the LLM inferred the correct option by 1) choosing the option with the most repeated expressions and 2) eliminating options that were self-contradictory.

We point out two problems in the generation process: First, generating four different incorrect options from one correct option became problematic, as the correct option naturally included more repeated words than the incorrect options. Second, not providing image and context information for option generation led to contradictory or incompatible expressions in some options. Therefore, from this experiment, we can conclude that to fairly evaluate reasoning ability, the process of generating choices should be improved to avoid providing additional information that could serve as a shortcut.

4.3 Good Option and Bad Option

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. But before improving the choice generation process, this question must be answered first: What distinguishes a good choice from a bad choice?

As we examined the augmented choice examples generated by the LLM, we were able to categorize the choices into three levels of quality, as shown in Figure 7. 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 questioning the change to orange can be considered the best type of choice. Next, choices that were possible to predict from the input image 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 and choices 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.

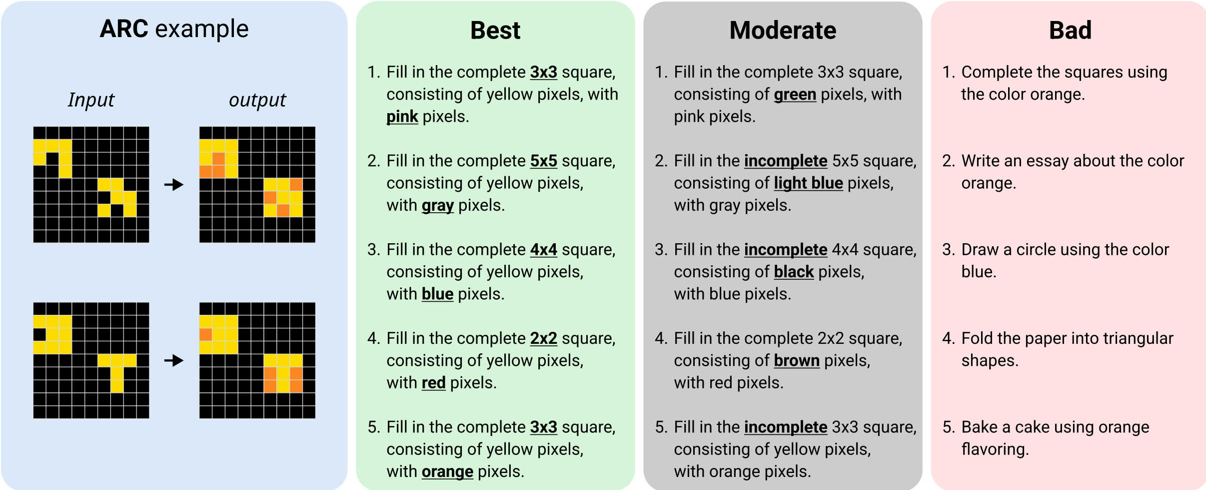

Figure 7: Three examples of multi-choice options augmented differently by the LLM. The given problem is to fill in an object with holes with the color orange to make a 3×3 square, where the size of the square and the color are the core aspects of the problem. The good example demonstrates an understanding 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.

5 Discussion

5.1 Limitations in the Multi-Choice Generation Method

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.

5.2 Limitations 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.

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 supplementary 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 propose specific conditions for designing multiple-choice questions that effectively evaluate the required reasoning abilities without enabling shortcuts.

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 paves the way for the development of more sophisticated and automated high-quality question generators.

7 Limitation

Our study has two main limitations. First, the generated options lack quality, allowing LLMs to find shortcuts. Second, there is a lack of metrics to measure the quality of the options. We have found issues such as repeated words and contradictory content in the current multiple-choice options. 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.

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. 1967. The Recognition Problem. *M.: Nauka*.

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*.

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. *arXiv:2404.13591*.

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*.

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*.

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*.

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.

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*.