Comparing Humans, GPT-4, and Claude 2 On Abstraction and Reasoning Tasks

Christian Bustos[1,*], Fridoon Najafi[1], Daniele Pennisi[1], and Karim Hamdar[1]

[1]Univerisity of Padova, Padova, Italy

Email addresses: davidechristian.mancosubustos@studenti.unipd.it (C. Bustos)*, fridoon.najafi@studenti.unipd.it (F. Najafi), daniele.pennisi@studenti.unipd.it (D. Pennisi), karim.hamdar@studenti.unipd.it (K. Hamdar) *Corresponding author, a, Padova, Italy

ABSTRACT

This study investigates the abstract reasoning capabilities of large pre-trained language models (LLMs), specifically GPT-4 and the focus of our experiment Claude 2, using the ConceptARC dataset. This dataset comprises analogy puzzles designed to test a solver's abstract reasoning skills. Our research reveals a notable performance disparity between these AI models and human solvers, with humans consistently outperforming the AI systems. This outcome underscores the significance of advanced prompting techniques in enhancing AI performance, indicating that the manner in which problems are presented markedly influences an AI's conceptual comprehension. Our experimental results suggest that Claude 2 not only exhibits subpar abstraction abilities compared to human levels but also underperforms relative to the abstraction capabilities previously demonstrated by GPT-4 in similar studies.

Keywords: Analogical Reasoning, ConceptARC, Enhanced Prompting, GPT-4, Claude 2

1 Introduction

The quest to imbue machines with abstract reasoning abilities comparable to human intelligence has been a subject of profound exploration in artificial intelligence. The advent of large pre-trained language models (LLMs) such as GPT-4 and Claude 2 has sparked intriguing debates about the emergence of abstract reasoning capabilities within these sophisticated systems. Abstract reasoning, characterized by the induction of rules or patterns from limited data and their application to novel scenarios, is a hallmark of human intelligence (manifested even in young children[1]). Recent assertions by researchers posit that sufficiently large pre-trained language models, exemplified by GPT-4, exhibit emergent abilities for abstract reasoning, general pattern recognition, and analogy-making[234]. However, contrary findings indicate that while LLMs excel at problem-solving within the confines of their training data, their generalization capabilities outside these parameters are limited[567]. Some interpretations suggest that LLMs lean towards learning intricate patterns of associations and resort to "approximate retrieval" rather than deploying generalizable abstract reasoning[8]. Given that proficiency in creating and reasoning with abstract representations is pivotal for robust generalization, a comprehensive understanding of how LLMs, such as Claude 2, have attained these abilities becomes imperative. This paper explores Claude 2's performance on ConceptARC[9] tasks, a set of analogy puzzles designed to assess general abstract reasoning capabilities. Through rigorous experimentation, we aim to shed light on the nuanced landscape of abstract reasoning in artificial intelligence, drawing comparisons between humans, GPT-4, and Claude 2.

The ConceptARC dataset[10], a variant of the original Abstraction and Reasoning Corpus (ARC)[11], presents a series of visual and analogical reasoning tasks that challenge humans and AI models to discern underlying patterns and apply them to novel situations. A crucial aspect of our study involves using enhanced prompts, a modification from the traditionally employed basic prompt. Our research pivots on the hypothesis that these enhanced prompts, which offer more context and guidance, could significantly influence the models' performance by providing explicit task directives and bridging the gap between human and AI analogical reasoning capabilities.

The intersection of these large language models' capabilities with the challenges posed by the ConceptARC dataset presents a unique opportunity to scrutinize the current landscape of abstract reasoning in AI. This exploration is not just a comparison of problem-solving skills but a deeper inquiry into the models' ability to emulate the intricate and nuanced process of human analogical thinking. By focusing on the comparative performance of Claude 2 and GPT-4 on this specific dataset and examining the impact of enhanced prompting, this study seeks to contribute a significant piece to the puzzle of AI's evolving cognitive abilities. The findings and insights from this research are poised to inform the future trajectory of AI development, particularly in cognitive modeling and the pursuit of machines capable of human-like abstraction and reasoning.

2 Methodology

In this section we delve into previous studies of abstract reasoning abilities across human participants and GPT-4 and describe the procedure used to obtain our experimetal results with Claude 2. The section is subdivided into two main parts: Material, describing the previous researches on which our work aim to expand; Prompting and Experiment Procedure, outlining the specific considerations and adaptations made to the experimental setup.

2.1 Material

2.1.1 The Abstraction and Reasoning Corpus

Chollet proposed the Abstraction and Reasoning Corpus (ARC)[12] as a benchmark to have a defined procedure that allows to evaluate such abilities in both humans and machines.

ARC consists of 1,000 manually created few-shots analogy puzzles of various complexity, each of which contains usually between 2 and 4 demonstrations (in the form of input-output grids) for each pattern, evaluating the solver ability through the relative input of the "test grid" that the solver has to transform using the inferred rule and obtaining the correct output (a series of examples is shown in the Figure 1). Chollet published 800 ARC tasks and kept the remaining 200 as a "hidden" test set[11].

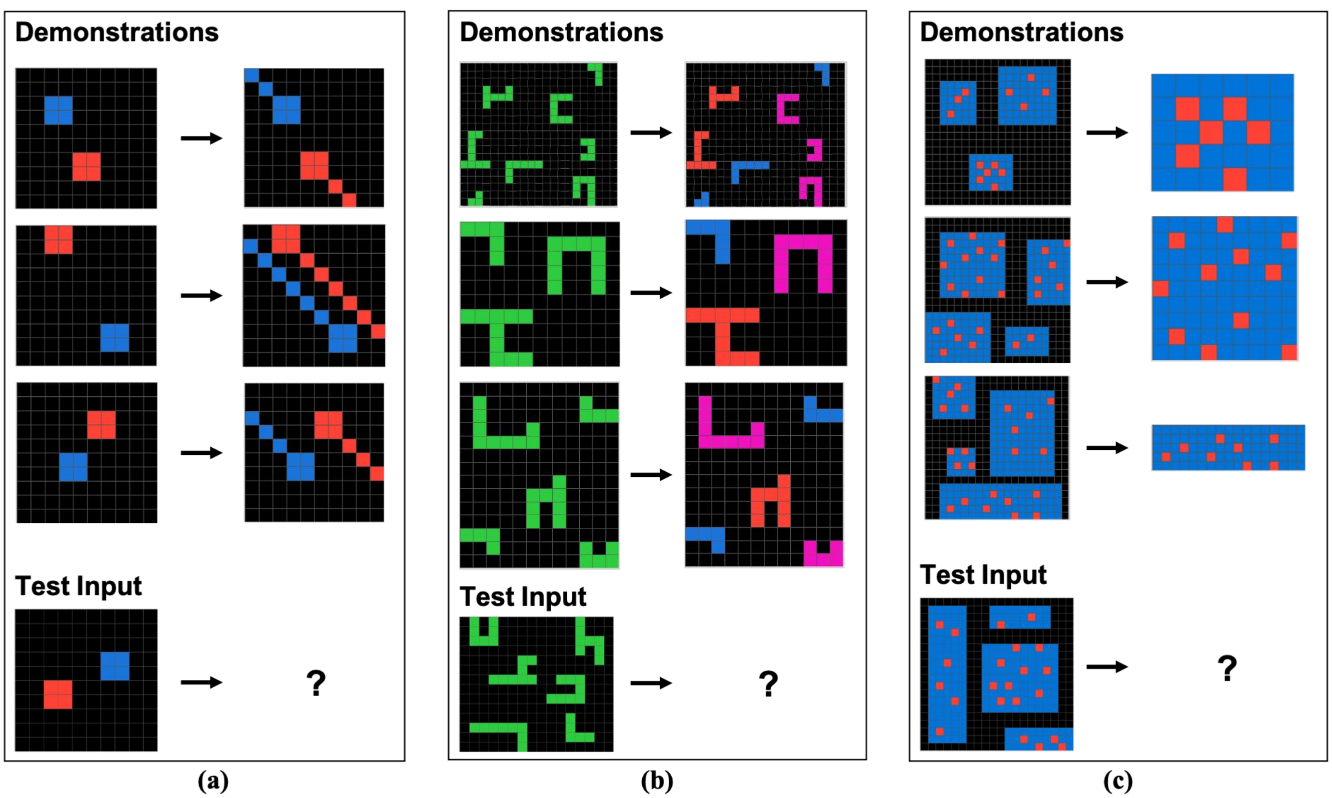


Figure 1. Examples of ARC tasks[11], with 3 different types of pattern (a), (b) and (c). [color should be used for this figure in print]

When constructing the dataset Chollet intentionally avoided the use of language and other "learned symbols" (that could have led to "approximate retrieval" and pattern matching based on prior training data), allowing to capture the core of abstract reasoning: inducing general rules or patterns from the training examples and applying these to previously unseen instances in a flexible way.

2.1.2 ConceptARC Conception and Experiments

Moskvichev et al.[9] detected two major problems in the original ARC corpus:

- Many of the tasks are quite difficult, even for humans, so it may be too high of a wall for AI systems to overcome.
- ARC does not offer systematic evaluation of understanding of particular core concepts, allowing to obtain only a general evaluation and not the specific strengths/weaknesses in the various underlying concepts.

To deal with these problems, the authors created a new benchmark in the ARC domain, ConceptARC[10], whose tasks are intentionally designed to be easy for humans. The dataset is composed by 48 minimal tasks (much simpler than the others) and by 480 standard tasks used for the actual evaluation, these are organized in 16 tipologies of systematic variations of particular core spatial and semantic concepts. Each concept group contains 30 tasks (and 3 minimal tasks), each of which instantiates the concept in a different way, and with differing degrees of abstraction. The author's claim was that high performance over this new dataset would be a better benchmark to evaluate the understanding of, and ability to reason abstractly about, the underlying concept.

Moskvichev et al. gave these tasks to human participants on the Amazon Mechanical Turk and Prolific platforms, and also tested GPT-4 on all 480 tasks (following the format shown in Figure 2). They found that human performance substantially exceeded that of GPT-4 and other systems; in particular, the overall accuracy of humans was 91%, while GPT-4 reached an accuracy of 19% and 25%, respectively using temperature 0 and 0.5, allowing to make the conclusion that even with the simplified benchmark the tested models lack the ability to generalize over the variations in a concept group, and thus failing to develop the analogical capabilities that ARC is meant to test.

2.1.3 Experiments with GPT-4 Using Enhanced Prompting

Mitchell et al.[13] noted that Moskvichev et al.'s evalation of GPT-4 had one important limitation: the prompt format that they used was overly simple and might not have communicated enough about the task. To address this limitation the authors evaluated GPT-4 on the ConceptARC benchmark using a much more expressive prompt that includes both instructions and an example of a solved task (pattern), allowing the model to contextualize the request.

The final prompt, that can be seen in Figure 3, was used through OpenAI's API on all 480 ConceptARC tasks, first with GPT-4's temperature set to 0 and then with temperature set to 0.5, to test the effects of temperature on performance.

The authors concluded that the more detailed prompting method resulted in a higher accuracy overall, 0.33 for both temperature settings, significantly exceeding the performances obtained with the basic prompt of 0.19 for temperature 0 and 0.25 for temperature 0.5. However they also concluded that "GPT-4's performance remains well below the high performance of humans, supporting the conclusion that, even with more informative prompting, the system lacks basic abstract reasoning abilities tested by this corpus".

2.2 Prompting and Experiment Procedure

Through this paper we tested the analogical capabilities of the AI model developed by Anthropic Claude 2, tested through the official website[14] and making a separate conversation for each task.

The starting prompt considered was the same used by Mitchell et al. (Figure 3), this was adapted and changed through various empirical tests applying slight changes (e.g. removing phrases, adding conceptual indications, changing the structure of the prompt, ecc.) made a subset of the minimal and standard tasks, to find the best performing solution. A crucial flaw that slowed the testing procedure (and also the completion of the experiment) and reduced its effectiveness was the limited amount of daily usages allowed by the Claude model (bort normal and premium versions).

The exact prompt used to do the final evaluation is reported in the Figure 4 (with the relative execution on Claude in the Figure 5). This prompt reached the best performances by applying the following changes to the Mitchell et al.'s prompt:

- Removal of the solved task, due to the tendency of the model to being mislead by the different pattern used as an example.

- Removal of the specific clues about the type of trasformation rule, keeping only the general ones about the approach, due to to the tendency of the model to being mislead by these indications.
- Insertion of the direct request immediatly before the input grid to be transformed, due to the tendency of the model to allucinate generating highly incorrect answers otherwise.

When providing the queries to Claude the response always included a textual explanation of the reasoning (often contrasting with the actual output grid provided) that due to it not being part of the interest of this study was ignored, considering only the actual output grid as the valid reply. If the reply contained an incorrect output grid, the request was repeated up to two times by using the "Retry" option offered in the Claude 2 website interface, for it to supply a different answer, for a maximum of three guesses, which is standard for all ARC anc ConceptARC evaluations. If a correct answer was generated within these three guesses, the task was considered to be solved.

3 Results

In this section we evalueted the analogical reasoning abilities of Claude 2 in the ConceptARC tasks, comparing its performance to human benchmarks and various configurations of the GPT-4 model mentioned in the previous sections. The results delineate a clear hierarchy in analogical reasoning across different concepts. The Table 1 summarizes the accuracies of humans, GPT-4 at different temperatures, and Claude 2 on ConceptARC tasks.

Concept	Humans	GPT-4 T=0	GPT-4 T=0.5	Claude 2	base GPT-4 T=0	base GPT-4 T=0.5
Above and Below	0.90	0.50	0.47	0.43	0.23	0.37
Center	0.94	0.37	0.37	0.33	0.33	0.33
Clean Up	0.97	0.43	0.46	0.13	0.20	0.27
Complete Shape	0.85	0.47	0.40	0.27	0.23	0.23
Copy	0.94	0.37	0.33	0.17	0.23	0.27
Count	0.88	0.27	0.23	0.30	0.13	0.17
Extend To Boundary	0.93	0.20	0.20	0.10	0.07	0.10
Extract Objects	0.86	0.13	0.13	0.07	0.03	0.07
Filled and Not Filled	0.96	0.27	0.30	0.20	0.17	0.27
Horizontal and Vertical	0.91	0.33	0.37	0.17	0.27	0.33
Inside and Outside	0.91	0.30	0.33	0.20	0.10	0.17
Move To Boundary	0.91	0.23	0.17	0.07	0.20	0.20
Order	0.83	0.27	0.30	0.07	0.27	0.27
Same and Different	0.88	0.23	0.30	0.27	0.17	0.27
Top and Bottom 2D	0.95	0.60	0.63	0.40	0.23	0.37
Top and Bottom 3D	0.93	0.30	0.27	0.13	0.20	0.27
All concepts	0.91	0.33	0.33	0.20	0.19	0.25

Table 1. Accuracies of humans, GPT-4, and Claude AI on ConceptARC tasks. "T" represents the temperature used for GPT-4, "base" indicates the use of the basic promp. The data for base GPT-4 and Humans was taken from the repository[10], the data for GPT-4 from the Mitchell et al.'s paper[13].

As observed in Table 1 humans exhibited superior performance with an average accuracy of 0.91, highlighting their natural proficiency in analogical reasoning. Claude 2's best performance was seen in "Above and Below" and "Top and Bottom 2D" with an accuracy of 0.43 and 0.40 respectively, closely followed by "Count" with 0.30, and by "Complete Shape"" and "Same and Different" at 0.27. However, its performance was notably weaker in categories like "Extract Objects", "Move to Boundary" and "Order", all with an accuracy of 0.07, indicating significant lack of understanding of the underlying concepts in these areas.

Compared to GPT-4 models, Claude 2 showed a disappointing level of performance. While it sligthly outperformed the base GPT-4 T=0 configuration, it was outpaced by all the other configurations. The results also evidentiate the difference between the two models: while Claude 2 seems to be more adept in "Count" tasks (in which it exceeded all the other GPT-4 configurations), it also severely lacks in comparison to GPT-4 in concepts like "Copy", "Horizontal and Vertical", "Top and Bottom 3D" and in particular "Move to Boundary" and "Order".

The difference of performance from humans to GPT-4 models, and then to Claude 2, emphasizes the challenge of achieving human-like analogical reasoning in AI systems. While Claude 2 demonstrates some hints of analogical reasoning it still falls short of the more nuanced human cognition.

4 Discussion

The results of this study offer insightful revelations about the analogical reasoning capabilities of LLMs, particularly highlighting the nuanced influence of enhanced prompting on performance. Notably, Claude 2's performance, when subjected to the improved prompting technique, underscores a mixed outcome, attesting the potential of Claude 2 to capture and process complex patterns but also showing the limitations in its current abilities to apply analogical reasoning consistently across a diverse spectrum of tasks, not reaching the performances of enhanced prompt's aided GPT-4.

The comparison with human performance further contextualizes these findings. Humans consistently outperformed both Claude 2 and GPT-4 across all conceptual categories, reinforcing the understanding that human cognition, with its inherent flexibility and depth, remains a lofty benchmark that AI has yet to meet. The performance gap elucidated in this study is not merely a quantitative measure but a qualitative indicator of the current state of AI in the context of high-level cognitive functions.

This research contributes to the broader discourse on AI and cognitive science by pinpointing the potential and pitfalls of current LLMs in analogical reasoning tasks. The nuanced impact of enhanced prompting on AI performance, as evidenced in this study, adds a layer of complexity to our understanding of AI's cognitive capabilities. It prompts a reconsideration of the design and application of prompting strategies and underscores the necessity for more sophisticated, context-aware, and adaptive approaches in future AI models.

In conclusion, this comparative study on the performance of Claude 2 and GPT-4 using the ConceptARC dataset reveals the intricate dynamics of AI's analogical reasoning capabilities and the influential role of enhanced prompting. While strides have been made, the journey towards achieving a human-like abstraction and generalization in AI is ongoing, marked by promising advancements and formidable challenges. The insights garnered from this research are instrumental in steering future explorations and innovations in pursuing more intelligent, wise, and cognitively robust AI systems.

References

1. Walker, C. M. & Gopnik, A. Toddlers infer higher-order relational principles in causal learning. *Psychol. science* 25, 161–169 (2014).
2. Wei, J. *et al.* Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).
3. Mirchandani, S. *et al.* Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721* (2023).
4. Webb, T., Holyoak, K. J. & Lu, H. Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* 7, 1526–1541 (2023).
5. McCoy, R. T., Yao, S., Friedman, D., Hardy, M. & Griffiths, T. L. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638* (2023).
6. Razeghi, Y., Logan IV, R. L., Gardner, M. & Singh, S. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 840–854 (2022).
7. Wu, Z. *et al.* Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477* (2023).
8. Kambhampati, S. Can llms really reason and plan. *Commun. ACM* (2023).
9. Moskvichev, A., Odouard, V. V. & Mitchell, M. The conceptARC benchmark: Evaluating understanding and generalization in the ARC domain. *arXiv preprint arXiv:2305.07141* (2023).
10. Odouard, V. V. ConceptARC repository. https://github.com/victorvikram/ConceptARC (2023).

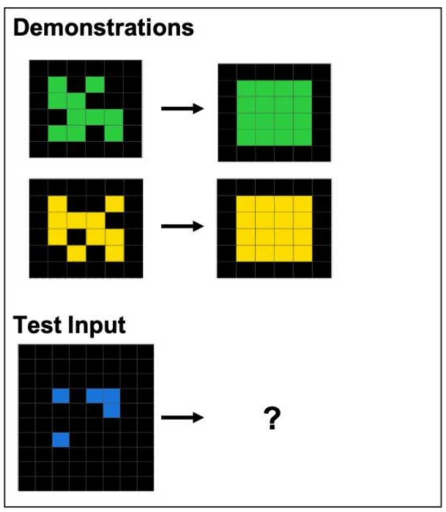
11. Chollet, F. The Abstraction and Reasoning Corpus (ARC) repository. https://github.com/fchollet/ARC (2019).
12. Chollet, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547* (2019).
13. Mitchell, M., Palmarini, A. B. & Moskvichev, A. Comparing humans, GPT-4, and GPT-4V on abstraction and reasoning tasks. *arXiv preprint arXiv:2311.09247* (2023).
14. Anthropic. Claude 2 model. https://claude.ai/ (2023).

5 Appendix

5.1 Prompts for GPT-4

In their evaluation of GPT-4, Moskvichev et al.[10] translated ConceptARC tasks into text representations used in basic prompts like the one shown in Figure 2, encoding the 10 possible colors as integers, and each row of a grid is encoded as a list of integers inside square brackets.

To better capture the abilities of GPT-4, Mitchell et al.[13] developed an enanched prompt using the same encoding, obtaining the result shown in Figure 3, that respects the format required by the OpenAI API.



(a)

System: "You are a helpful assistant that solves analogy making puzzles. Only give the answer, no other words or text."

User: "Let's try to complete the pattern: "

input 1: [0 0 0 0 0 0] [0 3 0 3 0 0] [0 3 3 0 0 0] [0 0 3 3 3 0] [0 3 3 0 3 0] [0 0 0 0 0 0]

output 1: [0 0 0 0 0 0] [0 3 3 3 3 0] [0 3 3 3 3 0] [0 3 3 3 3 0] [0 3 3 3 3 0] [0 0 0 0 0 0]

input 2: [0 0 0 0 0 0] [0 4 0 0 4 0] [0 4 4 4 0 0] [0 4 0 4 4 0] [0 0 4 0 4 0] [0 0 0 0 0 0]

output 2: [0 0 0 0 0 0] [0 4 4 4 4 0] [0 4 4 4 4 0] [0 4 4 4 4 0] [0 4 4 4 4 0] [0 0 0 0 0 0]

input 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 1 0 1 1 0 0] [0 0 0 0 0 1 0 0] [0 0 0 0 0 0 0 0] [0 0 1 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 0 0]

output 3:

(b)

Figure 2. (a) A visual task from the ConceptARC corpus. (b) The corresponding textual prompt used to obtain the GPT-4 evaluation[9]. [color should be used for this figure in print]

5.2 Prompts for Claude 2

Starting from the prompt reported at Figure 3 and following the same encoding we developed a simpler yet still enhanced prompt, due to the propensity of Claude of being misled. The structure of the final prompt obtained can be seen on Figure 4, with the respective execution through the Claude 2 model[14] on Figure 5.

```
# GENERAL INSTRUCTIONS
[System]
You will be given a list of input-output pairs  labeled "Case 0"  "Case 1"  and so on.
Each input and output is a grid of numbers representing a visual grid. There is a SINGLE rule that transforms each input
grid to the corresponding output grid.

The pattern may involve counting or sorting objects (e.g. sorting by size)  comparing numbers (e.g. which shape or symbol
appears the most? Which is the largest object? Which objects are the same size?)  or repeating a pattern for a fixed number
 of time.

There are other concepts that may be relevant.
- Lines  rectangular shapes
- Symmetries rotations  translations.
- Shape upscaling or downscaling  elastic distortions.
- Containing / being contained / being inside or outside of a perimeter.
- Drawing lines  connecting points  orthogonal projections.
- Copying  repeating objects.

You should treat cells with 0 as empty cells (backgrounds).

Please generate the Output grid that corresponds to the last given Input grid  using the transformation rule you induced
from the previous input-output pairs.

# EXAMPLE SOLVED TASK
[User]
Case 0:
Input:
[0 0 0 0 0 0] [0 2 2 0 0 0] [0 2 2 0 0 0] [0 0 0 0 0 0] [0 0 0 0 0 0] [0 0 0 0 0 0]
Output:
[0 0 0 0 0 0] [0 0 2 2 0 0] [0 0 2 2 0 0] [0 0 0 0 0 0] [0 0 0 0 0 0] [0 0 0 0 0 0]

Case 1:
Input:
[0 0 0 0 0 0] [0 0 4 4 0 0] [0 0 4 4 0 0] [0 0 0 0 0 0] [0 0 0 0 0 0] [0 0 0 0 0 0]
Output:
[0 0 0 0 0 0] [0 0 0 4 4 0] [0 0 0 4 4 0] [0 0 0 0 0 0] [0 0 0 0 0 0] [0 0 0 0 0 0]

Case 2:
Input:
[0 0 0 0 0 0] [0 3 0 0 0 0] [0 3 3 3 0 0] [0 0 0 3 0 0] [0 0 0 0 0 0] [0 0 0 0 0 0]
Output:
[0 0 0 0 0 0] [0 0 3 0 0 0] [0 0 3 3 3 0] [0 0 0 0 3 0] [0 0 0 0 0 0] [0 0 0 0 0 0]

Case 3:
Input:
[0 0 0 0 0 0] [0 0 0 0 0 0] [0 0 6 0 0 0] [0 0 6 0 0 0] [0 0 6 6 0 0] [0 0 0 0 0 0]

[Assistant]
Output:
[0 0 0 0 0 0] [0 0 0 0 0 0] [0 0 0 6 0 0] [0 0 0 6 0 0] [0 0 0 6 6 0] [0 0 0 0 0 0]

# QUERY TASK
[User]
Case 0:
Input:
[0 0 0 0 0 0] [0 3 0 3 0 0] [0 3 3 0 0 0] [0 0 3 3 3 0] [0 3 3 0 3 0] [0 0 0 0 0 0]
Output:
[0 0 0 0 0 0] [0 3 3 3 3 0] [0 3 3 3 3 0] [0 3 3 3 3 0] [0 3 3 3 3 0] [0 0 0 0 0 0]

Case 1:
Input:
[0 0 0 0 0 0] [0 4 0 0 4 0] [0 4 4 4 0 0] [0 4 0 4 4 0] [0 0 4 0 4 0] [0 0 0 0 0 0]
Output:
[0 0 0 0 0 0] [0 4 4 4 4 0] [0 4 4 4 4 0] [0 4 4 4 4 0] [0 4 4 4 4 0] [0 0 0 0 0 0]

Case 2:
Input:
[0 0 0 0 0 0] [0 6 0 0 6 6] [0 6 0 6 0 0] [0 6 0 6 0 0] [0 0 6 0 6 6] [0 0 6 6 6 0]

[Assistant]
Output:

# IF THE SYSTEM RETURNS WRONG ANSWER (REPEAT UP TO TWO TIMES)
[User]
Your answer does not solve the puzzle. Try again.

[Assistant]
I apologize for my mistake. Here is a better answer:
Output:
```

Figure 3. Example of the prompt used to test GPT-4 on ConceptARC tasks. The symbol "#" indicates comments not given in the actual prompt.

```
[User]
You will be given a list of Input-Output pairs labeled "- Case 1", "- Case 2" and so on.
Each Input and Output is a grid of numbers represeting a visual grid. There is a SINGLE rule that transforms each Input grid
to the corresponding Output grid.

You should treat cells with 0 as empty cells (backgrounds).

Here follows the Inputs-Output pairs:

# QUERY TASK EXAMPLE
- Case 1
Input:
[0,0,0,0,0,0],[0,0,2,2,0,0],[0,0,2,2,0,0],[0,0,0,0,0,0],[6,6,6,6,6,6],[0,0,0,0,3,3],[0,0,0,0,3,3],[0,0,0,0,0,0],[0,0,0,0,0,0]
Output:
[2,2],[2,2]

- Case 2
Input:
[0,0,0,7,7,0],[0,0,0,7,7,0],[0,0,0,0,0,0],[0,0,0,0,0,0],[6,6,6,6,6,6],[0,0,0,0,0,0],[0,0,4,4,0,0],[0,0,4,4,0,0],[0,0,0,0,0,0]
Output:
[7,7],[7,7]

- Case 3
Input:
[0,0,0,0,0,0,0,0],[0,8,8,0,0,0,0,0],[0,8,8,0,0,0,0,0],[0,0,0,0,0,0,0,0],[6,6,6,6,6,6,6,6],[0,0,0,0,0,0,0,0],[0,0,0,0,0,9,9,0],
[0,0,0,0,0,9,9,0]
Output:
[8,8],[8,8]

Please generate the Output grid that corresponds to the following Input grid using the transformation rule you induced from
the previous Input-Output pairs.
Input:
[0,0,3,3,0,0],[0,0,3,3,0,0],[0,0,0,0,0,0],[6,6,6,6,6,6],[0,0,0,0,0,0],[2,2,0,0,0,0],[2,2,0,0,0,0],[0,0,0,0,0,0],[0,0,0,0,0,0]
Output:

[Assistant]
*Textual answer* # TO BE IGNORED
*Output grid*

# IF THE SYSTEM RETURNS A WRONG ANSWER (REPEAT UP TO 2 TIMES)
[User]
*Use of the Retry function*

[Assistant]
*Textual answer* # TO BE IGNORED
*Output grid*
```

Figure 4. Example of the prompt used to test Claude 2 on ConceptARC tasks. The symbols "#" and "*...*" indicate comments not given in the actual prompt.

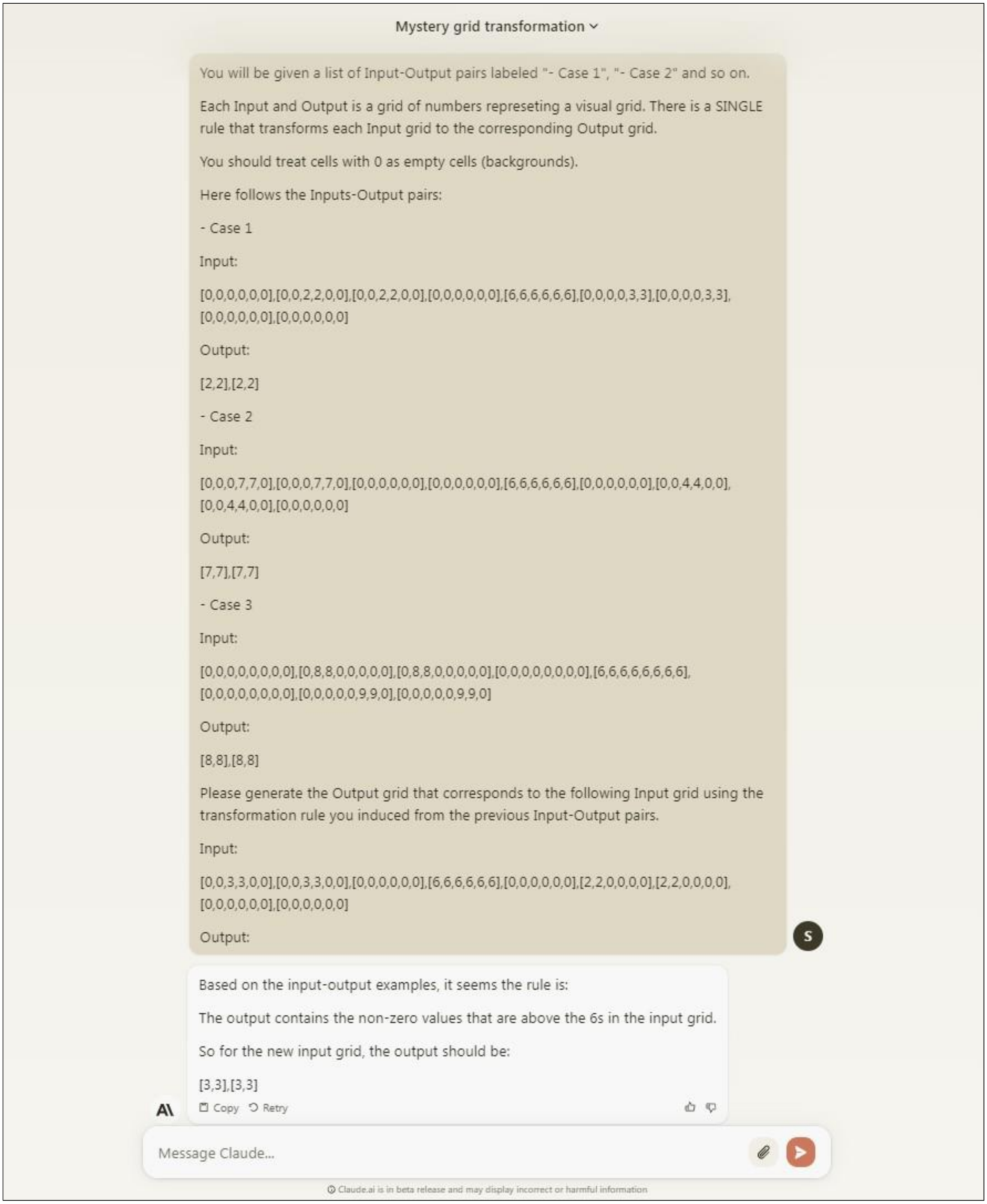


Figure 5. Example of execution through the Claude 2 interface[14].