# Claude 3's Performance on the ConceptARC Benchmark: An Analogical Reasoning Assessment

Davide Mazza, Thomas Rosso

**Abstract**—This study evaluates the abstract reasoning capabilities of Claude 3, a recently developed large language model, using the ConceptARC benchmark. Through systematic experimentation and enhanced prompting strategies, we assess Claude 3's performance across various analogical reasoning tasks within ConceptARC and compare it to previous results for humans and GPT-4. Our findings reveal a mixed picture, with Claude 3 exhibiting proficiency in certain conceptual areas while lagging behind in others. The study highlights the nuanced influence of prompting techniques on model performance and contributes insights to the broader pursuit of artificial intelligence systems capable of robust abstraction and reasoning. The discrepancies between Claude 3 and human-level performance underscore the ongoing challenges in achieving human-parity in abstract reasoning within AI.

## 1 Introduction

The quest to imbue artificial intelligence (AI) systems with human-like abstract reasoning abilities has been a longstanding pursuit in the field of cognitive science and AI research. Abstract reasoning, characterized by the capacity to discern patterns, induce rules, and apply them to novel scenarios, is a hallmark of human cognition [1]. Recent advancements in large language models (LLMs), such as GPT-4 and Claude, have ignited debates about their emergent capabilities in tackling complex abstract reasoning tasks. Previous studies have explored the performance of GPT-4 on the ConceptARC dataset [2], a benchmark designed to evaluate general abstract reasoning skills through a series of visual analogy puzzles. Mitchell et al. [3] investigated the impact of enhanced prompting techniques on GPT-4's performance, concluding that while more informative prompts improved the model's accuracy, its capabilities remained inferior to human-level performance. This finding supports the notion that achieving human-like abstract reasoning in AI systems remains a formidable challenge. The advent of Claude, a more recent LLM developed by Anthropic, presents an opportunity to further probe the frontiers of AI's abstract reasoning abilities. This study aims to evaluate Claude's performance on the ConceptARC dataset, drawing comparisons with the previously reported results for humans and GPT-4. By leveraging enhanced prompting strategies and rigorous experimentation, we seek to shed light on the nuanced landscape of analogical reasoning in AI, contributing insights to the broader discourse on cognitive modeling and AI development.

## 2 Methodology

In this section, we delve into the details of the ConceptARC dataset and the previous studies that evaluated human and GPT-4 performance. We also outline the prompting strategy and experimental procedure employed for assessing Claude's analogical reasoning capabilities.

### 2.1 Material

#### *2.1.1 The ConceptARC Dataset*

The ConceptARC dataset [2], derived from the original Abstraction and Reasoning Corpus

(ARC) [4], comprises a series of visual analogy puzzles designed to assess abstract reasoning skills systematically. The dataset consists of 528 tasks, including 48 minimal tasks and 480 standard tasks organized into 16 conceptual categories. Each category contains 30 tasks that instantiate a specific spatial or semantic concept, such as ”Above and Below,” ”Complete Shape,” or ”Extract Objects,” with varying degrees of abstraction.

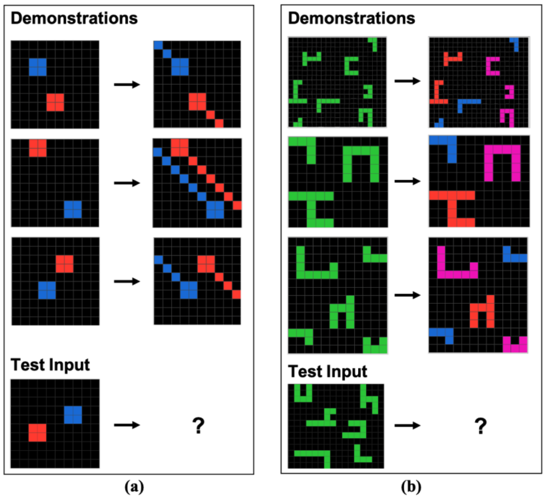


Fig. 1: Example of ARC Puzzle

Previous evaluations have established human performance as a benchmark, with an average accuracy of 91% across all ConceptARC tasks [2]. Additionally, Moskvichev et al. [2] and Mitchell et al. [3] evaluated GPT-4's performance using basic and enhanced prompting techniques, respectively, reporting accuracies ranging from 19% to 33%.

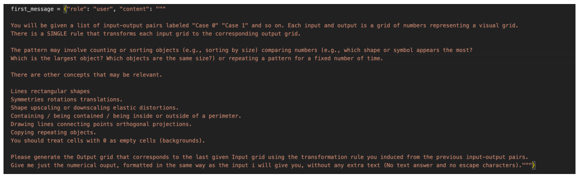
### 2.1.2 Prompting Strategies for GPT-4

Mitchell et al. [3] recognized the limitations of the basic prompting approach used by Moskvichev et al. [2] and developed an enhanced prompt for GPT-4. This prompt included detailed instructions and an example of a solved task, providing more context and guidance to the model. The enhanced prompting strategy resulted in improved performance, with GPT-4 achieving an accuracy of 33% on the ConceptARC tasks, significantly higher than the 19% accuracy obtained with basic prompting.

## 2.2 Prompting and Experiment Procedure for Claude

Building upon the insights from Mitchell et al. [3], we employed a dynamic approach to develop an effective prompting strategy for Claude. Starting with the enhanced prompt used for GPT-4, we introduced modifications and variations through empirical testing on a subset of ConceptARC tasks. Particularly we decided to develop a Python notebook capable to interact with the API of Claude 3, being able to automate the process of asking the AI to solve the puzzle, check the answer and eventually ask to give other answers (in the case the previous were wrong) and finally save the results.

More specifically, the problem is first introduced to Claude 3, using a similar message to the one used in the Mitchell et. al paper [3]. With a slight modification of their message the problem is kindly and cleared explained so the AI can understand the messages that will arrive later. This prompt aimed to strike a balance between providing sufficient context and guidance while avoiding potential sources of confusion or misleading information.The introductory prompt is visible in Figure 2.



```
first_message = {"role": "user", "content": """

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.
Give me just the numerical ouput, formatted in the same way as the input i will give you, without any extra text (No text answer and no escape characters)."""}
```

Fig. 2: Instruction message given to Claude

Now that the AI has a full context of what is the task and how will happen, it is now time to process the different JSON files in the Concept ARC Repository. Each one is analysed, slightly modified and converted to string to conform to Claude 3's API; the test part in the current JSON file is removed except for the first one (without its output) and the full string is sent to Claude 3 to be processed. The answer of Claude 3 is preprocessed and put in the same format as the one in the JSON file; the two are then compared. In this case, if the answer is correct the result is saved and we pass to the next test;

instead if the answer is wrong we ask Claude 3 to try again. Similarly to Mitchell et. al, the total number of attempts are 3. If after the third attempt the answer is still wrong we pass to the next test. This full run procedure is then repeated for every test and file in the Concept ARC Repository (see Next Section).

### 2.3 Limitations and Problems during the run of the procedure

Since Claude 3 usage is limited by daily usage quota and the API are usable under payment, not all the files could be analyzed leading to a biased evaluation process. To make the code useful and more fair, we decided to use the Minimal JSON for every test in the Repository and in addition to this three other files picked at random in every folder. Using this technique we were able to obtain 12 test results per category (3 from the minimal and 9 more from the tests picked at random) keeping the costs of using the APIs relatively low and being able to cover all the categories. Obviously this is not sufficient to obtain some fair results and being able to compare the obtained data with the one in the Mitchell et. al paper (they had 30 results per category); these results will be discussed in the Results section.

## 3 RESULTS

This section presents the results of our evaluation of Claude 3's performance on the ConceptARC benchmark. The results are compared with human performance and previous evaluations of GPT-4. The analysis focuses on the model's accuracy across various conceptual categories, highlighting areas of strength and weakness. We also discuss the impact of different prompting strategies on the model's performance. The table below summarizes the comparative accuracy of humans, GPT-4 (at two temperature settings), and Claude 3 across the 16 conceptual categories within the ConceptARC dataset.

### 3.1 Results adjustment

As said in the Section 2.3, these obtained results from Claude 3 are not sufficient to make a fair comparison between the outcomes from ChatGPT and Humans; however they can provide us an initial look at the capabilities of Claude 3 Model in the task of Analogical Reasoning. Since obtaining more results from Claude was too expensive (in terms of money) we initially thought about using a scale factor to our results (in order to put them in the same range as the ones in the Mitchell et. al Paper) but after some trials we have not found a multiplication unit fair enough to provide meaningful results. In the Table 1 it is possible to observe the obtained results per category.

### 3.2 Results comment

Overall, Claude 3 achieved an average score of 0.55 across all conceptual categories, which is significantly lower than the human average of 0.91 but higher than the GPT-4 averages of 0.33 for both temperatures. This suggests that while Claude 3 is still far from human performance, it shows improvement over GPT-4 in these tasks.

In the ″Clean Up″ category, Claude 3 scored 0.75, approaching the human score of 0.97 and significantly outperforming GPT-4 (0.43 and 0.46). This indicates that Claude 3 demonstrates a strong performance in this area. Similarly, in the ″Complete Shape″ category, Claude 3's score of 0.83 is comparable to the human score of 0.85 and significantly higher than GPT-4's scores of 0.47 and 0.40, indicating near-human competence.

On the other hand, in categories like ″Order″ and ″Move To Boundary,″ Claude 3's scores are significantly lower. For instance, in ″Order,″ Claude 3 scored 0.17, which is much lower than the human score of 0.83 and slightly lower than GPT-4 (0.27 and 0.30). In ″Move To Boundary,″ Claude 3's score of 0.17 is also much lower than the human score of 0.91, indicating a clear area of weakness.

Other categories show mixed results. In ″Above and Below″ and ″Center,″ Claude 3 scored 0.58, which is lower than the human scores (0.90 and 0.94 respectively) but higher than GPT-4's scores (0.50 and 0.47 for ″Above and Below,″ and 0.37 for ″Center″). This indicates a reasonable understanding but still leaves room for improvement.

The ”Extend To Boundary” category shows Claude 3 scoring 0.67, which is lower than the human score of 0.93 but significantly better than GPT-4’s 0.20, suggesting a marked improvement. Similarly, in ”Top and Bottom 2D,” Claude 3 scored 0.83, close to the human score of 0.95 and higher than GPT-4 (0.60 and 0.63), demonstrating solid performance.

Conversely, in ”Extract Objects,” Claude 3 scored 0.50, which is much higher than GPT-4’s 0.13 but still lower than the human score of 0.86. In the ”Same and Different” category, Claude 3 scored 0.33, which is below the human score of 0.88 but comparable to GPT-4 (0.23 and 0.30).

## 4 Conclusions

The evaluation of Claude 3’s performance on the ConceptARC benchmark provides valuable insights into the current state of AI in the realm of analogical reasoning. Despite the inherent biases in our results, primarily due to the limited number of test samples per category compared to the comprehensive dataset used in previous studies, several key conclusions can be drawn. Claude 3 shows a notable improvement over GPT-4 in nearly all conceptual categories. This suggests that the advancements in Claude 3’s architecture and the enhanced prompting strategies we employed contribute to better performance in analogical reasoning tasks. The improvements are particularly evident in categories such as ”Clean Up” and ”Complete Shape,” where Claude 3’s scores approach those of human participants. However, despite these improvements, Claude 3 still falls short of human-level performance in every category. The average score of 0.55 for Claude 3, compared to the human average of 0.91, underscores the ongoing challenges in developing AI systems that can match human abstract reasoning capabilities. Categories like ”Order” and ”Move To Boundary” highlight significant areas of weakness for Claude 3, indicating the need for further refinement and enhancement. The limited number of test samples per category in our study (12 compared to 30 in the Mitchell et al. study) introduces a bias that affects the generalizability of our results. This limitation emphasizes the necessity for future studies to use larger, more representative datasets to provide a more accurate assessment of AI performance. The results indicate specific areas where Claude 3 excels and others where it struggles. Future research should focus on addressing the weaknesses identified, particularly in categories where Claude 3’s performance significantly lags behind human benchmarks. Additionally, employing more extensive and varied datasets will be crucial in obtaining a comprehensive evaluation of Claude 3’s capabilities. Our findings also highlight the significant impact that enhanced prompting strategies can have on AI performance. The tailored prompts used in this study helped improve Claude 3’s accuracy, suggesting that further refinement of these techniques could yield even better results. While Claude 3 demonstrates promising advancements over previous models like GPT-4, there remains a considerable gap to human-level abstract reasoning. Addressing the biases in our study and continuing to refine both the AI models and the evaluation methodologies will be essential steps towards achieving more robust and reliable AI systems capable of human-like reasoning.

## References

[1] A. Walker, C. M. Gopnik, “Toddlers infer higher-order relational principles in causal learning.” *Psychol. science 25, 161–169*, 2014.

[2] O. V. V. . M. M. Moskvichev, A., “The conceptarc benchmark: Evaluating understanding and generalization in the arc domain.” *arXiv preprint arXiv:2305.07141*, 2023.

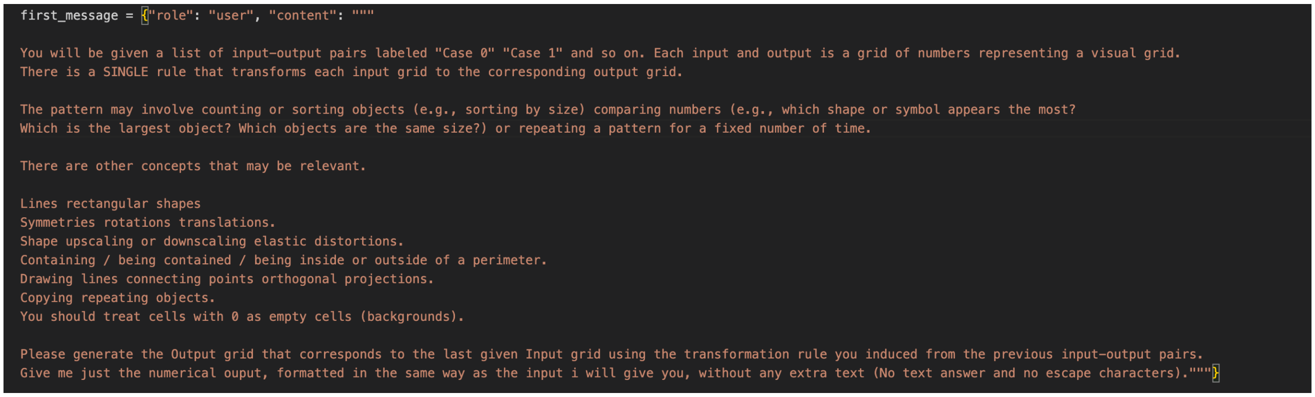
[3] P. A. B. . M. A. Mitchell, M., “Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks.” *arXiv preprint arXiv:2311.09247*, 2023.

[4] F. Chollet. (2019) The abstraction and reasoning corpus (arc) repository. [Online]. Available: https://github.com/fchollet/ARC(2019).

| Concept | Humans | **GPT-4 Temp = 0** | **GPT-4 Temp = 0.5** | **Claude 3** |
|---|---|---|---|---|
| Above and Below | 0.90 | 0.50 | 0.47 | 0.58 |
| Center | 0.94 | 0.37 | 0.37 | 0.58 |
| Clean Up | 0.97 | 0.43 | 0.46 | 0.75 |
| Complete Shape | 0.85 | 0.47 | 0.40 | 0.83 |
| Copy | 0.94 | 0.37 | 0.33 | 0.75 |
| Count | 0.88 | 0.27 | 0.23 | 0.33 |
| Extend To Boundary | 0.93 | 0.20 | 0.20 | 0.67 |
| Extract Objects | 0.86 | 0.13 | 0.13 | 0.50 |
| Filled and Not Filled | 0.96 | 0.27 | 0.30 | 0.58 |
| Horizontal and Vertical | 0.91 | 0.33 | 0.37 | 0.50 |
| Inside and Outside | 0.91 | 0.30 | 0.33 | 0.75 |
| Move To Boundary | 0.91 | 0.23 | 0.17 | 0.17 |
| Order | 0.83 | 0.27 | 0.30 | 0.17 |
| Same and Different | 0.88 | 0.23 | 0.30 | 0.33 |
| Top and Bottom 2D | 0.95 | 0.60 | 0.63 | 0.83 |
| Top and Bottom 3D | 0.93 | 0.30 | 0.27 | 0.50 |
| **All concepts** | **0.91** | **0.33** | **0.33** | **0.55** |

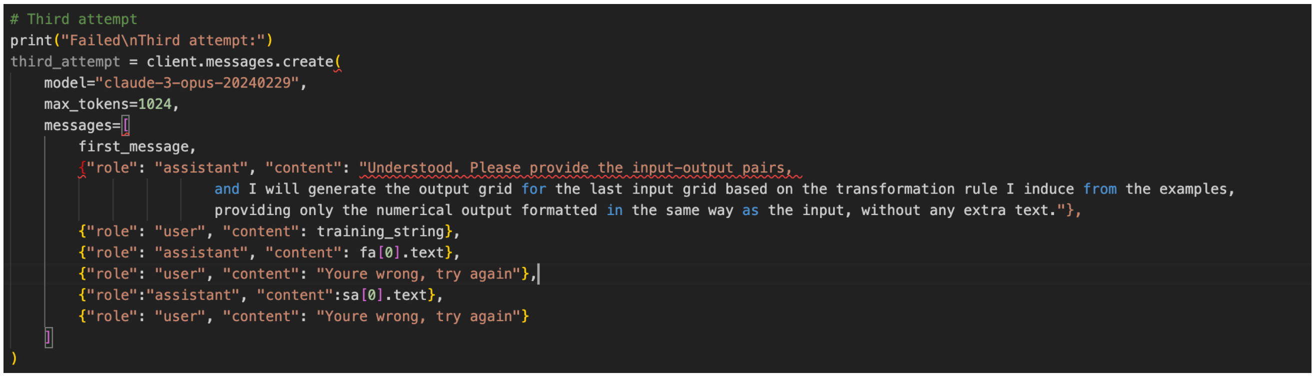
TABLE 1: Comparison of Concept Understanding Between Humans, GPT-4 at Different Temperatures (0 and 0.5), and Claude 3 across Various Conceptual Categories within the ConceptARC Benchmark.

# APPENDIX A
# PROMPTS GIVEN TO CLAUDE 3 AND SOME RESULTS FROM THE CODE



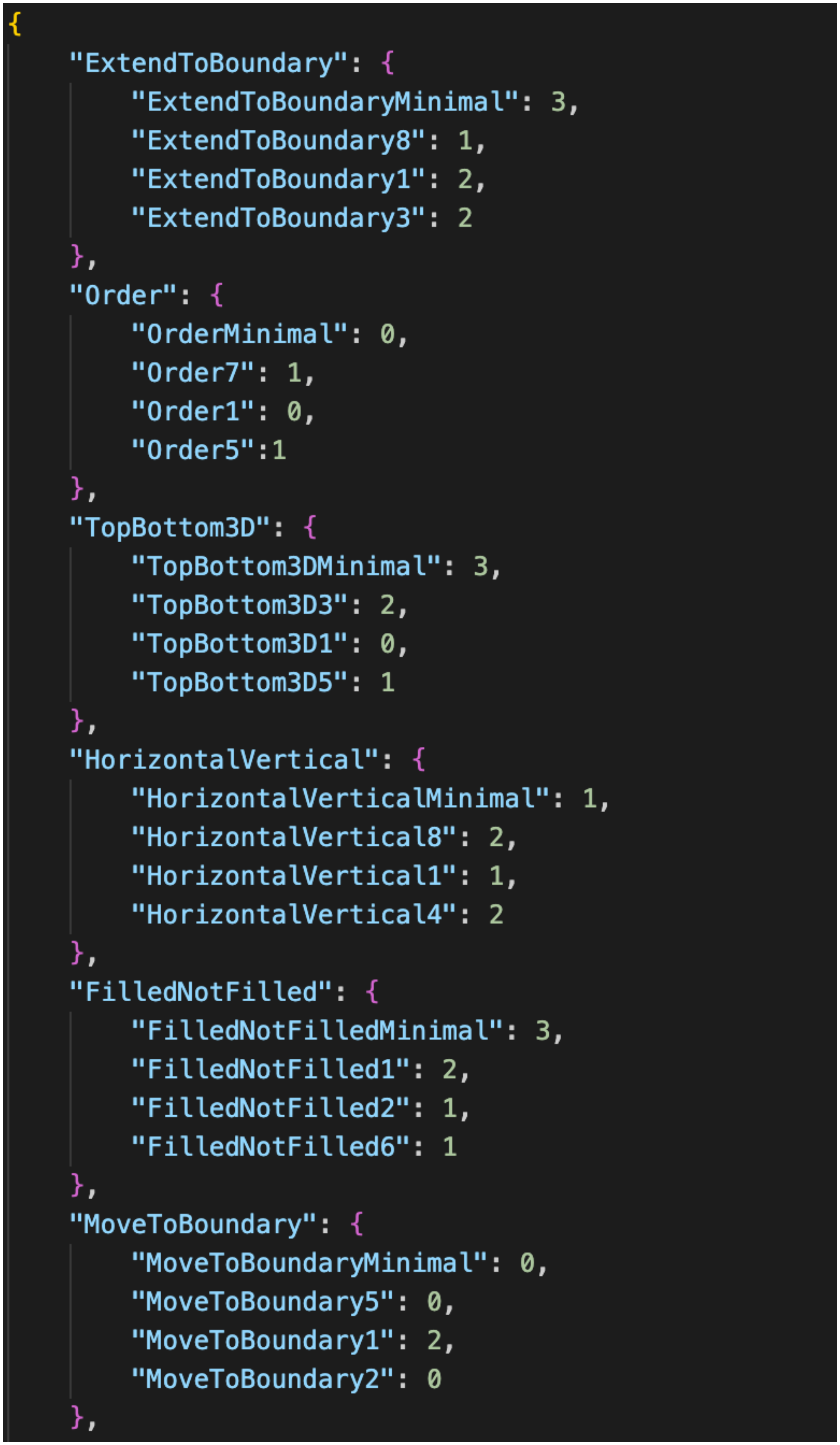
```
first_message = {"role": "user", "content": """

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.
Give me just the numerical ouput, formatted in the same way as the input i will give you, without any extra text (No text answer and no escape characters)."""}
```

Fig. 3: Instruction message given to Claude



```python
# Third attempt
print("Failed\nThird attempt:")
third_attempt = client.messages.create(
    model="claude-3-opus-20240229",
    max_tokens=1024,
    messages=[
        first_message,
        {"role": "assistant", "content": "Understood. Please provide the input-output pairs,
                        and I will generate the output grid for the last input grid based on the transformation rule I induce from the examples,
                        providing only the numerical output formatted in the same way as the input, without any extra text."},
        {"role": "user", "content": training_string},
        {"role": "assistant", "content": fa[0].text},
        {"role": "user", "content": "Youre wrong, try again"},
        {"role":"assistant", "content":sa[0].text},
        {"role": "user", "content": "Youre wrong, try again"}
    ]
)
```

Fig. 4: Message given to Claude in the case of multiple mistakes (Note that the code is wrong on purpose just to fit the text in the image)



```
{
    "ExtendToBoundary": {
        "ExtendToBoundaryMinimal": 3,
        "ExtendToBoundary8": 1,
        "ExtendToBoundary1": 2,
        "ExtendToBoundary3": 2
    },
    "Order": {
        "OrderMinimal": 0,
        "Order7": 1,
        "Order1": 0,
        "Order5":1
    },
    "TopBottom3D": {
        "TopBottom3DMinimal": 3,
        "TopBottom3D3": 2,
        "TopBottom3D1": 0,
        "TopBottom3D5": 1
    },
    "HorizontalVertical": {
        "HorizontalVerticalMinimal": 1,
        "HorizontalVertical8": 2,
        "HorizontalVertical1": 1,
        "HorizontalVertical4": 2
    },
    "FilledNotFilled": {
        "FilledNotFilledMinimal": 3,
        "FilledNotFilled1": 2,
        "FilledNotFilled2": 1,
        "FilledNotFilled6": 1
    },
    "MoveToBoundary": {
        "MoveToBoundaryMinimal": 0,
        "MoveToBoundary5": 0,
        "MoveToBoundary1": 2,
        "MoveToBoundary2": 0
    },
```

Fig. 5: Some of the results obtained from the code, every subindex of the relative category indicates the score on base 3 (e.g. OrderMinimal = 0 means that it scored 0 out of 3).