Comparing Humans, Llama-2-13B, On Abstraction and Reasoning Tasks

Parisa Foroozande Nejad, Ramtin Reyhani Kivi, Ehsan Eslami Shafigh

January 21, 2024

Abstract

We evaluated the abstract reasoning abilities of Llama-2-13B using the ConceptARC benchmark [1], a tool designed to assess the comprehension and application of core knowledge concepts. Our findings support the conclusion of a previous study [2] (which analyzed different versions of GPT-4) that Llama-2-13B also lacks strong abstraction abilities comparable to humans.

1 Introduction

In this report we investigate the extent to which large pre-trained language models (LLMs) have developed the ability to reason abstractly. Abstract reasoning is the ability to extract rules or patterns from limited data and apply them to new situations, a key aspect of human intelligence that even young children can develop from a few examples. [3]

Some researchers have suggested that LLMs can develop emergent abilities for reasoning [4], pattern recognition, and analogy-making [5]. However, the underlying mechanisms of these abilities are not fully understood, and others have questioned whether LLMs truly form human-like abstractions [6]. These skeptics argue that LLMs may instead rely on learning complex patterns from their training data and applying them in new situations through "approximate retrieval." [7]

The ability to create and reason with abstract representations is fundamental to robust generalization, making it crucial to evaluate LLMs' progress in this area. This study assesses Llama-2-13B's performance on ConceptARC tasks, a collection of analogy puzzles that test general abstract reasoning abilities.

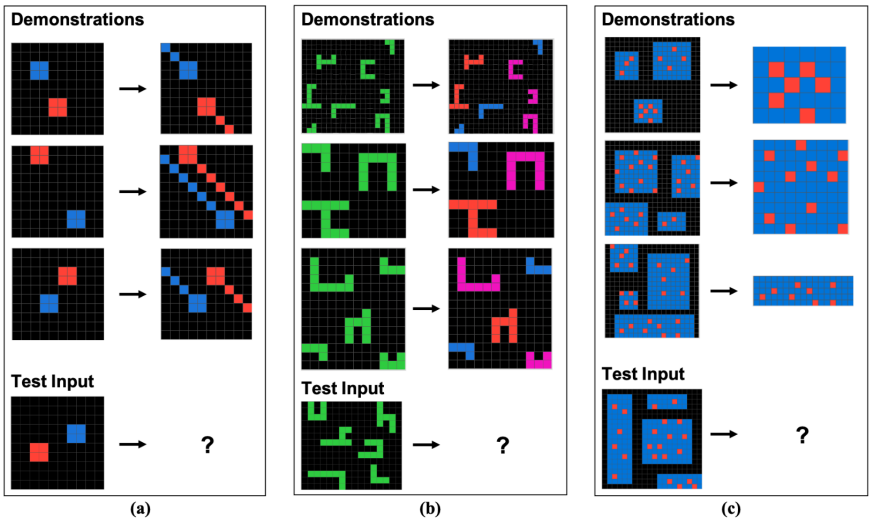


Figure 1: Examples of ARC tasks from [8]. Each task has a set of demonstration input-output pairs that illustrate an abstract grid-transformation rule, and a test input. The solver's challenge is to generate a new grid that results from applying the abstract rule to the test input. (Figure is from [1])

2 The Abstraction and Reasoning Corpus

Abstraction and Reasoning Corpus (ARC) is a benchmark for evaluating the ability of both humans and machines to solve analogy puzzles involving grids. It consists of 1,000 manually created tasks, each of which contains a small number of demonstration transformations. The solver must induce the abstract rule underlying the demonstrations and apply it to a test grid to generate a transformed output grid. Figure 1 gives three examples of ARC tasks. [9]

ARC was designed to test a subset of the core knowledge systems hypothesized to be innate in humans, including objectness, numerosity, and basic geometry and topology. Notably, it omits knowledge of language or other learned symbols, which helps avoid the potential for "approximate retrieval" and pattern matching based on prior training data. Instead, ARC aims to capture the essence of abstract reasoning: inducing general rules or patterns from small numbers of examples and applying them flexibly to new situations.

3 Previous Evaluation of LLMs using the ConceptARC Benchmark

The original ARC corpus has two main drawbacks. First, it includes some very difficult tasks that may hinder the advancement of AI systems capable of reasoning in this domain. Second, it lacks a systematic evaluation of the understanding of fundamental concepts. This means that even if a system can solve an individual ARC task, it doesn't necessarily imply a thorough comprehension of the underlying concepts.

To address these issues, Moskvichev et al. [1] created a new benchmark called ConceptARC, specifically designed to be easily understandable for humans. It comprises 480 tasks that systematically vary specific core spatial and semantic concepts like top, bottom, inside, outside, center, and same/different. Each concept group contains 30 tasks, each representing the concept differently and varying in abstraction level. The authors believe that high performance across these diverse instantiations of a concept indicates a solid understanding and ability to abstractly reason about it.

Moskvichev et al. tested these tasks on human participants and two winning programs from the Kaggle ARC challenge, along with GPT-4. They discovered that human performance significantly outperformed machines across all concept groups in the corpus. Human accuracy was 0.91, compared to 0.52 for the first-place Kaggle program and 0.25 for GPT-4 (using temperature 0.5). The authors concluded that humans' consistently high accuracy on each concept suggests successful generalization across different variations within each concept group. In contrast, the significantly lower accuracies of the programs indicate a lack of ability to generalize and a failure to develop the abstractions that ARC aims to assess.

Moskvichev et al.'s evaluation of GPT-4 had a key limitation: the prompt format they used was overly simplistic and might not have conveyed enough information about the tasks. To address these problems Mitchell et al. [2] used a more detailed prompt. The prompt provided instructions and an example of a solved task. If GPT-4 responded incorrectly, it was asked to provide a different answer up to three times. This is standard practice for ARC evaluations [10]. If a correct answer was given within three tries, the task was considered solved. GPT-4 achieved an overall accuracy of 0.33 using the one-shot prompting method, compared to 0.19 for the zero-shot method. This indicates that the one-shot method provides more information to GPT-4, leading to improved performance. However, GPT-4's performance is still significantly below human performance, which is around 0.90. This suggests that even with more informative prompting, GPT-4 lacks the basic abstract reasoning capabilities tested by ConceptARC.

In this work we tested Llama-2-13B with ConceptARC benchmark by applying the same methodology used by Mitchell et al.

Concept	Human Accuracy	Llama 2-13B Accuracy	Llama 2-13B Dimension Accuracy	GPT-4 Text-Only Accuracy
Above and Below	0.90	0.03	0.2	0.47
Center	0.94	0.13	0.23	0.37
Clean Up	0.97	0.00	0.27	0.46
Complete Shape	0.85	0.00	0.23	0.40
Copy	0.94	0.00	0.13	0.33
Count	0.88	0.07	0.2	0.23
Extend to Boundary	0.93	0.00	0.1	0.20
Extract Objects	0.86	0.00	0.07	0.13
Filled and Not Filled	0.96	0.03	0.17	0.30
Horizontal and Vertical	0.91	0.00	0.1	0.37
Inside and Outside	0.91	0.00	0.2	0.33
Move to Boundary	0.91	0.00	0.07	0.17
Order	0.83	0.10	0.23	0.30
Same and Different	0.88	0.03	0.3	0.30
Top and Bottom 2D	0.95	0.00	0	0.63
Top and Bottom 3D	0.93	0.00	0.23	0.27
All concepts	0.91	0.03	0.17	0.33

Figure 2: Accuracies of humans, Llama-2-13B and GPT-4 (with temperature 0.5) on each concept group (30 tasks) and over all concepts (480 tasks) in ConceptARC, using the prompt given in the appendix. The results on humans are from [1]The results for GPT-4 are from [2]. The "Dimension Accuracy" shows the relevance of the answers returned by the model with respect to the ground truth answers.

4 Experiments Evaluating Llama-2-13B on ConceptARC Tasks

To evaluate Llama-2-13B, we adapted the prompting from mitchel et al. [2] This prompt provides detailed instructions about the task and an example of a solved task. If Llama-2-13B responds with an incorrect answer, we ask it to provide a different answer, up to three times. If a correct answer is generated within these three tries, the task is considered solved.

Using this prompting method, we tested Llama-2-13B on all 480 ConceptARC tasks (30 per each of the 16 concept groups). We set Llama's temperature to 0.5. The accuracies (fraction of solved tasks within each concept group and overall) are shown in Figure 2, along with human accuracies from another study[1], and the results from mitchel et al [2] for text-only version of GPT-4.

Since Llama-2-13B accuracy was not promising, we introduced a new measure "Dimension Accuracy" to evaluate the relevance of the answers returned

by the model with respect to the ground truth answers. We calculate this accuracy to check if the dimension of the given answer is the same dimension of the correct answer.

5 Conclusion

This study evaluated the abstract reasoning performance of Llama-2-13B using the ConceptARC corpus, a tool designed to measure abstract reasoning skills through core concepts. The research method was similar to that used by Mitchel et al. to evaluate abstract reasoning in GPT-4. Our results showed that Llama-2-13B performed significantly poorer than GPT-4 on all ConceptARC corpus tasks. This outcome was expected, as smaller language models like Llama-2-13B may not have the capacity for abstract reasoning, which is a complex skill that emerges with increasing model size.

Since we used 1-shot prompting and doing so we reached the maximum prompt length possible with Llama-2, we believe the observed results represent the best possible performance of Llama-2-13B on the ConceptARC tasks translated into text representations.

References

[1] A. Moskvichev, V. V. Odouard, and M. Mitchell. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *Transactions On Machine Learning Research*, 2023.

[2] Melanie Mitchell and Alessandro B. Palmarini and Arseny Moskvichev. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. 2023.

[3] Caren M. Walker and Alison Gopnik. Toddlers infer higher-order relational principles in causal learning. *Psychological Science*, 25(1):161–169, 2014. ISSN 09567976, 14679280. URL http://www.jstor.org/stable/24539479.

[4] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. *arXiv e-prints*, art. arXiv:2206.07682, June 2022. doi: 10.48550/arXiv.2206.07682.

[5] Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models, 2023.

[6] Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. Large language models are not strong abstract reasoners, 2024.

[7] Subbarao Kambhampati. Can llms really reason and plan? 2023.

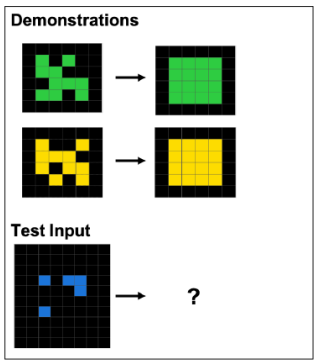
[8] F. Chollet. The abstraction and reasoning corpus (arc). 2023.

[9] François Chollet. On the measure of intelligence, 2019.

[10] Kaggle.com. Kaggle abstraction and reasoning challenge. 2020. Accessed 2023-11-09.

6 Appendix

Following Moskvichev et al. [1] to feed the ConceptARC tasks into our model, we translate the tasks into text representations, like the example shown in Figure 3.



(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 3: (a) A task from the ConceptARC corpus. (b) The corresponding prompt used in [1] to give to GPT-4. (Image is from [1]; best viewed in color.)

Based on the Llama2 structure we divided the prompt into the general prompt (Figure 4) and the system prompt (Figure 5). The prompt consists of a list of input-output pairs. The first pair is the example solved task, and the second pair is the query task. (Figure 6) The system prompt provides general instructions and rules for completing the task. The temperature parameter defines the randomness of the outputs. In this project we set the temperature to 0.5.

In case the answer returned by the model is wrong for a test, we prompt the model with the text below:
"[INST] The last output you produced is wrong, try again by predicting another output for the last input given. [/INST]"
We try this up to 2 times. Note that we need to keep the number of tokens below 4096, which limits the number of trials for each test.

```
    output_generator = replicate.run(
      "meta/llama-2-13b-chat:f4e2de70d66816a838a89eeeb621910adffb0dd0baba3976c96980970978018d",
      input={
          "prompt": prompt,
          "system_prompt": system_prompt,
          "max_new_tokens": 2048,
          "temperature": 0.5
      }
  )
```

Figure 4: The general prompt for Llama-2-13B using the Replicate API that provides the basic inputs shown in the figure.

```
system_prompt = """
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. Just give the output with no further words.
"""
```

Figure 5: The system prompt provides general instructions and rules for completing the task.

Figure 6: An example of the prompt used to test Llama-2-13B on ConceptARC tasks. The "[INST]" and "[/INST]" notations are used for instructions given to Llama-2-13B. The prompt consists of a complete solved example and another part for testing.