When is it enough?
Investigating Pythia Models' ability to complete induction tasks

Samuele Antonelli, Simone Baratella, Kassandra Briola
CIMeC, University of Trento
`(samuele.antonelli | simone.baratella | kassandralea.briola)@studenti.unitn.it`

Link for the project contents: `https://rb.gy/dn7yhb`.

1 Introduction

Large Language Models (LLMs) show good abilities in a lot of tasks derived from cognitive psychology and decision making (Binz Schultz, 2023; Lampinen et al., 2022). Yet, sometimes their answers are far away from what a human would give, being too precise and resulting unnatural, even when the request is very general. Therefore, our focus lies in when these language models start to provide human-like answers and when those answers go beyond human-likeness. We're specifically interested in inductive reasoning tasks. In order to investigate this phenomenon, we are leveraging the Pythia Models (Biederman et al., 2023) and humans, comparing their answers to prompts created ad hoc.

2 Background

2.1 Inductive Reasoning

Induction is a reasoning task largely used in psychology and cognitive sciences where, starting from a set of specific information, the subject is required to understand the connection and to create a novel case. It is often described as a reasoning method that allows one to draw conclusions going from particular to general, opposite to a deduction task (Murphy et al., 2010; Feeney et al., 2007) .
Many psychology experiments had exploited this concept to study the field of reasoning and decision making (Hayes et al., 2010).

2.2 Pythia Models

Pythia Models is an open source suite of 16 models provided by EleutherAI (Biederman et al., 2023), which allows to access various steps of training, along with the data used. All the models are trained on the same tokens (for a total amount of 299.9B) in order to avoid biases. Moreover, they are trained on different numbers of layers, dimensions and attention heads (check `https://github.com/EleutherAI/pythia?tab=readme-ov-file#using-pythia` for further information).

3 Hypotheses & Implications

H1: The models show a human-like performance beginning from a specific amount of parameters, that keep increasing as the number of parameters grows, until they starts worsening off and fall out of human-likeness range.
If H1 is true: Optimal human-like performance in LLMs is accomplished by staying in a defined range of parameters. Going over this range would be counterproductive.
H2: The models show a human-like performance, starting from a specific amount of parameters and don't show a following decrease.

***If H2 is true**:* Optimal human-like performance in LLMs is accomplished by having a defined amount of parameters. Added parameters won't worsen or better the inductive performance.
H0: The models never show a human-like performance.
***If H0 is true**:* The model used in this study is not fit for various reasons to accomplish the task.

4 Project Design

We created 36 original prompts that require to perform an induction task. This initial pool was proposed to human judges (three, different from who created them) and then brought down to 27, discarding the most misleading ones.
The prompts were then divided into three categories, in order to investigate both the inductive reasoning and the ability to move between and within categories during the task.
Firstly, the tasks were proposed to humans via Google Form, using the results to compute a score of human-likeness. Then, the same tasks were prompted to the models.
Both the answers of humans and models were analyzed through two main measures:
- **Accuracy:** which investigates how much the answers match the prompt.
- **Relative Similarity:** that examines the coherence within multiple answers (when provided).

4.1 Prompt Designing

To design our prompts, we first defined three categories in which to divide them: animals, objects and food.
Each prompt describes an animal, an object or a food, and can be distinctly inserted in one of the three categories. Each prompt has a structure comprehending 5 features, as example (1).

*(1) List up to three things with these features: **four legs, whiskers, paws, chase mice, purr.***

The prompts share the same main structure and are divided in sets of three, where two belong to the same category and the last belong to another one. Inside these sets, each prompt has three base-features and two defining features. Example (2) and (3) show two different sets of three prompts.

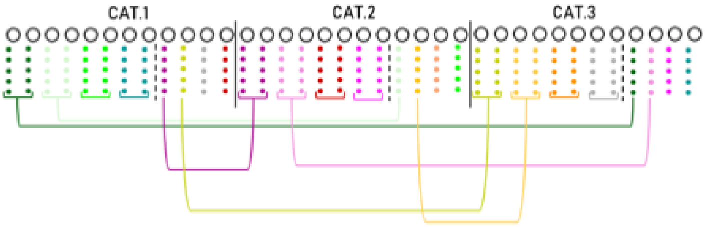
(2)
A. *List up to three things with these features: four legs, whiskers, paws, chase mice, purr.*
B. *List up to three things with these features: four legs, whiskers, paws, chew bones, obey orders.*
C. *List up to three things with these features: four legs, whiskers, paws, stuffed, soft.*

(3)
A. *List up to three things with these features: long, thin, flexible, strangle their prey, venomous.*
B. *List up to three things with these features: long, thin, flexible, can climb with, can do knots with.*
C. *List up to three things with these features: long, thin, flexible, covered in plastic, electricity passes through it.*

In (2), the three main features are kept the same, while the last two change moving within categories (animal to animal, *cat* to *dog*) and between categories (animal to object, *dog* to *peluche*).

The image (4) shows a visual example of the relationship between the 27 prompts:

(4)



We decided to ask up to three answers to explore the next prediction and consistency in the models and their counterpart in humans. We didn't ask a specific amount of answers in order to avoid forced ones, which wouldn't reflect a normal reasoning process.

For consulting the full list of prompt, please see the Appendices at chapter 9.

4.2 Human Data Gathering

We used a Google Form to gather data from humans. The prompts were shuffled and every answer was mandatory. We collected a total of 23 human answers.

4.2.1 Exclusion Criteria

To exclude outliers, we added three prompts to check the attention of participants. Those prompts were built as the classical ones, but at the end a specific answer was required.
Prompt (5) is an example:

(5) List up to three things with these features: made of metal, usually cylindrical, contains liquid, and is found in the kitchen. Attention check, write "garage."

The presence of such attention checks was also stated at the beginning of the form. People who failed the checks were excluded.

4.3 Models

4.3.1 Pyhtia Models and implementations

For this project we based our code on the one provided by EleutherAI, available at `https://github.com/EleutherAI/pythia`, and implemented some features to get better results:

Response handling function:
Built-in function needed to modulate the behavior of the model. We best optimized the model for our inductive task by setting the following variables:

- max_lenght = 100; for optimal response length (in tokens).
- temperature = 0.4; it controls the randomness in the response. Low value = grounded response.
- top_p = 0.2; nucleus sampling, controls the randomness of the prediction by focusing on the most likely next words. Lower = more varied output.
- top_k = 30; limits the pool of next word candidates. Only the top 30 most likely next words are considered for sampling, keeping the model on the topic even if top_p is low.
- no_repetition penalty=1.5; discourage the repetition of tokens to avoid repetition loops.
- no_repeat_ngram_size = 2; avoid repetition of word sequences, extra safety measure to ensure a wider exploration of the possible answer.
- num_beams = 2; the model keeps the 2 best candidates at each step and optimizes at the end for the highest overall probability.

Interaction loop:
A while loop that allows multiple interactions. It requests a string as an input and provides the model's answer. It allows multiple interactions without having to start the program, until explicitly required to exit.

For more detailed information, the code we used is available at `https://rb.gy/dn7yhb`.

5 Analyses

Once all the answers were collected we proceeded analyzing them.
The final score for a single model/human is given by a mean of the score for each prompt. While,

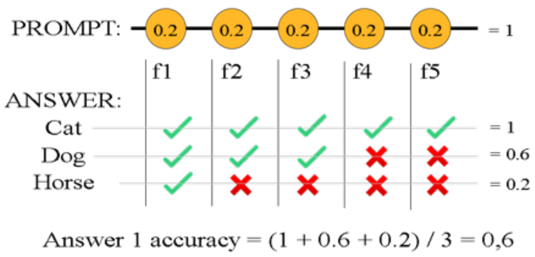
each prompt's score is given by the accuracy in case of a single answer, and the mean of accuracies added to relative similarity, in case of multiple answers.

5.1 Accuracy

Accuracy is the measure of how much the answers match the five features.
Each of the five features in each prompt was weighted 0.2. We then checked for each answer how many of them were met.
Example (6) shows the process: here the five features are all met in the first answer, two are missing in the second, and poorly met in the third. In this case the overall accuracy score is 0,6.

(6)

- **PROMPT:** List up to three things with these features: four legs, whiskers, paws, chase mice, purr.
- **FEATURES:** Four legs, whiskers, paws, chase mice, purr.
- **ANSWERS:** Cat, Dog, Horse.



We decided not to weigh the features differently among them, even though we acknowledge the fact that some characteristics may be more decisive in the understanding process.
We preferred to keep the calculation straight, and use the accuracy measure to see how many of the features were met in the answer, rather than a weighted score.
Our goal here is not to look at which feature brought to the conclusion, but rather to see how much models can resemble humans. If a feature allows humans to understand easier, it should show the same effect in human-like models.

5.2 Relative Similarity

We used relative similarity as a measure of how coherent the answers were among them. If we somehow created a misleading prompt, the answers should've been related in any case.
In order to measure it, we used human judgements, scoring individually each set of answers.

Relative similarity was divided into 4 sub-parts:

1. **Do the answers belong to the same category?** They all belong or not to one same category, such as food, animals, etc.
   Example:
   - *Apple and Cherry: both belong to foods.*
   - *Apple and Rope: belong to different categories (foods and objects).*

2. **Do the answers belong to the same subcategory?** They all belong or not to a subcategory inside of a bigger one.
   Example:
   - *Apple and Cherry: both foods and, inside foods, both fruits.*
   - *Apple and Bacon: both foods, but one is a fruit, the other is not.*

3. **Do the answers belong to the same type?** They are or aren't two types of the same thing, as well as hypernym/hyponym.
Example:
    - *Red apple and Yellow apple: both food, both fruits, are same type (apple).*
    - *Red apple and Cherry: both food, both fruits, but not the same type.*

4. **Do the answers belong to the same domain?** They belong to the same conceptual space, showing some type of connection in human judges' opinions.
Example:
    - *Red apple and Yellow apple: both food, both fruits, are same type, and can be found together (e.g. at the market).*
    - *Turkey and Fork: different categories, (food and objects), not the same subcategory, not the same type, but can be found together (on a dining table).*
    - *Red apple and Cat: not the same category, not in the same subcategory, not the same type and not in the same domain.*

The judgements were based on independent scores assigned by three different people. The scores were binary, each of the four sub-parts of relative similarity could either be judged as met= 1, or not met= 0. We decided to keep three independent measures to avoid biases, since different opinions might emerge.

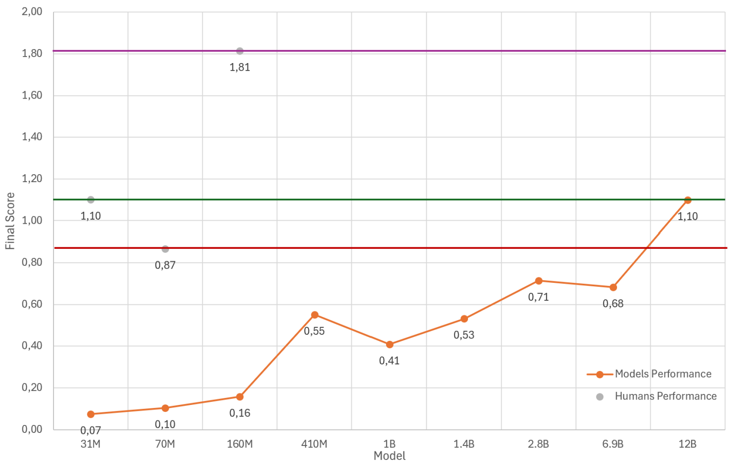
Lastly, the score for relative similarity was added to the accuracy one (in case of multiple answers).
Values were computed this way in order to avoid the unequal weighting of relative similarity in the scores of the ones who gave a single answer to each prompt (both models and humans).

6 Results

6.1 General Results

The human performance zone is shown in (7) by the red line for worst human final score, the violet line the best one, and the green one for the mean.
The models' performances are shown in orange.
Our results exhibit an ability to achieve a human-like performance only in the 12B model, which surpasses the lowest human performance and lays on humans' average score.

For consulting the tables with the answers and scoring, please see the Appendices at chapter 9.
(7)

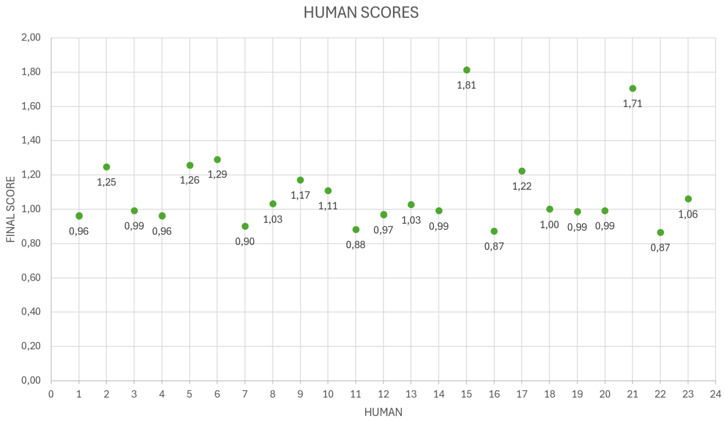


6.2 Human Scores

The image (8) looks at humans' scores individually. We can notice a range in performance that goes from a final score of 0,87 to 1,29. We also collected two outliers (participant 15 and participant 21) which scored higher. We kept the scores, since they don't show faulty reasoning or inability to complete the task, but just more effort in doing it, thus are still representative of human performance.

For consulting the tables with the answers and scoring, please see the Appendices at chapter 9.

(8)



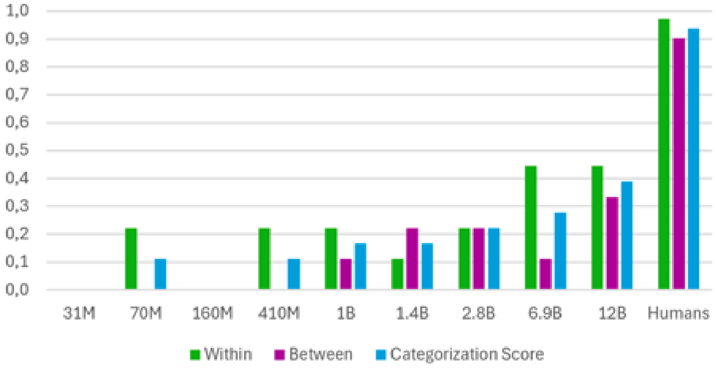
6.3 Categorization

Lastly, we plot the categorization in image (9). We wanted to check the models' capacity to move within the same category (green line) and between different categories (violet lines); blue lines show the mean of ability in categorization.

We can notice a huge gap in the performance between models and humans. The models that do not show any scoring were not able to complete at least one categorization correctly.

For consulting the tables with the answers and scoring, please see the Appendices at chapter 9.

(9)



7 Conclusion and Discussion

The obtained results do not allow us to draw any firm conclusions regarding the initial question, however, they do give us the opportunity to study the phenomenon in greater depth. Although H0 could be discarded, only the last model enters the realm of human-like performance. Indeed, the 12B model shows a final score that lays exactly on the mean of humans' performances. However, we don't know how the pace of the curve might change when using more complex models, whether it would just stay in the human range or leave it. Still, we can actually see that 12B parameters are sufficient to accomplish the proposed task well enough.

We noticed some peculiarities in our results: First, the fact that the curve doesn't show linear growth. Instead, the 410M model shows a first peek with a significant difference from the previous model (160M). Then, immediately afterwards, the 1B model's score worsens off, just as the 6.9B score is lower than 2.8B. It seems that, from the 410M model onwards, the growth increases and decreases in a cyclical repetition. Perhaps, a model with more than 12B parameters would fall outside the human range if it followed this pattern. It's possible to notice the biggest change in the answers in the 410M model. The smaller ones tend to give mostly nonsensical answers or clearly copy-pasted from a corpus of data. From the 410M model, more coherent responses are given. This change is reflected in Table (7) in a big jump, outperforming other models. We hypothesize that this amount of parameters might be an important turn-point in models' reasoning abilities. This peculiar behaviour could be due to the number of layers, dimensions and attention heads in each model. Indeed, consulting the Pythia documentation provided by EleutherAI (Biederman et al., 2023, `https://huggingface.co/EleutherAI/pythia-6.9b`), the 410M model is trained on 24 layers, 1024 dimensions and 16 attention heads, while the 1B goes back to 16 layers, 2048 dimensions and 8 attention heads. Lastly the 1.4B, which shows a performance similar to 410M is again trained on 24 layers and 16 attention heads.

Secondly, starting from the 2.8B model, we noticed that the models give more and more specific answers, not only just reasonable ones, highlighting how, as the number of parameters increases, the answers become more specific. For example the 410M model gives "bird" as an answer when describing a volatile in Antarctica, while the 12B gives "penguin". Similarly, the majority of people gave "penguin" as an answer, except for outliers who gave specific answers, such as "emperor penguin". Our data show how humans tend usually to generalize, but not to overgeneralize, and the same tendency seems to be shared by models. Initially, the models are only able to give hyper-general answers, which then become more grounded as the amount of parameters increases. We could not see the over-specification process, but hypothesize it would be the next step.

Lastly, we decided to investigate categorization during the induction task. The results plotted in Table (8) show how the humans perform much better than any given model. Moving within categories (e.g. from food to food) is easier for both models and humans than the passage between categories (e.g. from food to animal). These results are not surprising when we think about human psychology and phenomena like *priming*, where exposure to a stimulus influences the response to a subsequent, related stimulus (Schacter & Buckner, 1998). Still, it's interesting to look at how models never show a human-like performance, except the 12B, which still exhibits a large gap with humans. In this project the task is primarily designed to study induction. To investigate the phenomenon of categorization, a specific task would be needed.

8 Critical Evaluation

8.1 Limitations

We are conscious of some limitations in our work, which we're going to discuss in this section:

- **Time limit:** acknowledging the limitation in time given for this project, we were unable to test different promptings and implementations of the models. Therefore, we drew conclusions on the data and results gathered in the given temporal window.
- **Participants:** we involved mainly university students as participants. A possible limitation is the lack of English native speakers. The majority of participants do not have English as their mother tongue, and this might have led to some errors in understanding or reasoning.
- **Models:** we decided to use the Pythia models since they showed the best features for our aims. Still, there's room for discussion. Those models were built on up to 12B parameters, while the largest models available (such as GPT3 and GPT4) range from 175B to 1,76T parameters. We assume that disposing of bigger models might lead to one of our initial hypotheses.
  Also, the amount of training tokens might be important to take into account. Pythia models are trained on 299,9B tokens, while the gold-standard models are trained on an amount that ranges from 500B to 1,3T.
- **Fine-tuning:** the models we used were not fine-tuned for chatting. Especially in the smallest ones, we had to interpret the results, since they tended to give us sentences describing an object/food/animal, better than a list. Perhaps, training and fine-tuning the models for specific reasoning tasks would lead to better results.
- **Code:** we optimized the code with our current knowledge on coding. With more time and experimentation on it, we could have achieved a better interaction system.
- **Prompts:** prompts were created by us and judged by three independent judges, to state if they were appropriate and the features we found actually belonged to the target answer. A possible implementation would've been to use a corpus that allows us to pick words with the same frequency.

8.2 Future Directions

Lastly, we would like to give some possible future directions on how we would implement this work.

1. Use bigger models, trained on the same set of data, in order to check how the performance behaves after the 12B one.
2. Focus on which features actually allow the models and humans to specifically understand the target answer, more than on the ability to do so. Some line of research already exists on similar approaches, focusing on understanding how LLMs reason in order to come to a target answer (Bertolazzi et al., 2023; Han et al., 2024). Shifting the focus from large models, as ChatGPT, to more restricted ones, as Pythia, might be interesting.
3. Try a multilingual approach with native speaker. It would be interesting to use this same procedure with multiple languages and native speakers.
   Clearly the Pythia models are trained on American English. Yet, it's possible to dispose of the plain architecture without training. Thus, it would be possible to create similar experiments in different languages.

9 Appendices & References

9.1 Appendices

To get further information and full access to the data used, please check: `https://rb.gy/dn7yhb`. There you can find:

- The ReadMe file;
- The codes used for all the models;
- A PDF file, called *PROMPTS*, containing all the prompts created;
- **Two sheets** called *Grading Tables*, containing all the collected values from humans and models, all the scoring and the graphs computed as results.

9.2 Use of external resources

In this project the use of ChatGPT-4 was implemented to get an initial structure of the modification required to the code, and to create formulas for the Excel spreadsheets.
In no case was it used to generate content, write the final report or falsify the results. It was implemented only as a tool to facilitate the work pipeline.

9.3 Bibliography

[1] Bertolazzi, L., Mazzaccara, D., Merlo, F., Bernardi, R. (2023, September). ChatGPT's Information Seeking Strategy: Insights from the 20-Questions Game. In Proceedings of the 16thInternational Natural Language Generation Conference (pp. 153-162).

[2] Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., ... Van Der Wal, O. (2023, July). Pythia: A suite for analyzing large language models across training and scaling. In International Conference on Machine Learning (pp. 2397-2430). PMLR

[3] Binz, M., Schulz, E. (2023). Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences, 120(6), e2218523120.

[4] Feeney, A. E., Heit, E. E. (2007). Inductive reasoning: Experimental, developmental, and computational approaches. In Fifth International Conference on Thinking, Jul, 2004, University of Leuven, Belgium; Many of the chapter authors for this book talked at the aforementioned symposium.. Cambridge University Press.

[5] Han, S. J., Ransom, K. J., Perfors, A., Kemp, C. (2024). Inductive reasoning in humans and large language models. Cognitive Systems Research, 83, 101155.

[6] Lampinen, A. K., Dasgupta, I., Chan, S. C., Sheahan, H. R., Creswell, A., Kumaran, D., ... Hill, F. (2022). Language models show human-like content effects on reasoning tasks. arXiv preprint arXiv:2207.07051.

[7] Murphy, G. L., Ross, B. H. (2010). Category vs. Object Knowledge in Category-based Induction. Journal of memory and language, 63(1), 1–17. https://doi.org/10.1016/j.jml.2009.12.002

[8] Schacter, D. L., Buckner, R. L. (1998). Priming and the brain. Neuron, 20(2), 185-195.

[9] Wang, R., Zelikman, E., Poesia, G., Pu, Y., Haber, N., Goodman, N. D. (2023). Hypothesis search: Inductive reasoning with language models. arXiv preprint arXiv:2309.05660.