UC Merced
Proceedings of the Annual Meeting of the Cognitive Science Society

Title
Do large language models solve ARC visual analogies like people do?

Permalink
https://escholarship.org/uc/item/4bp4m6cf

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors
Opielka, Gustaw
Rosenbusch, Hannes
Vijverberg, Veerle P
et al.

Publication Date
2024

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Do large language models solve ARC visual analogies like people do?

Gustaw Opiełka (g.j.opielka@uva.nl)

Hannes Rosenbusch (h.rosenbusch@uva.nl)

Veerle Vijverberg (v.p.vijverberg@uva.nl)

Claire E. Stevenson (c.e.stevenson@uva.nl)
University of Amsterdam, Department of Psychological Methods, Amsterdam, Netherlands

Abstract

The Abstraction Reasoning Corpus (ARC) is a visual analogical reasoning test designed for humans and machines (Chollet, 2019). We compared human and large language model (LLM) performance on a new child-friendly set of ARC items. Results show that both children and adults outperform most LLMs on these tasks. Error analysis revealed a similar "fallback" solution strategy in LLMs and young children, where part of the analogy is simply copied. In addition, we found two other error types, one based on seemingly grasping key concepts (e.g., Inside-Outside) and the other based on simple combinations of analogy input matrices. On the whole, "concept" errors were more common in humans, and "matrix" errors were more common in LLMs. This study sheds new light on LLM reasoning ability and the extent to which we can use error analyses and comparisons with human development to understand *how* LLMs solve visual analogies.

Keywords: analogical reasoning; human vs AI cognition; large language models; abstract visual reasoning

Introduction

Until recently, visual analogy solving (e.g., ● is to ○ as ■ is to ?) was considered something that is easy for humans, but out of reach for AI deep learning models (Mitchell, 2021). However, large language models (LLMs) such as OpenAI's ChatGPT now appear capable of solving a range of analogy tasks in the text domain, including numeric and text-based versions of the matrix analogies in the Raven's Progressive Matrices (Webb et al., 2023; Hu et al., 2023) and, to a lesser degree, open-ended visual analogies (Moskvichev et al., 2023; Mitchell et al., 2023; Xu et al., 2023). The question then arises of *how* LLMs solve these visual analogies. Is the process similar to adult humans that identify abstract relations and map these to new instances? Or perhaps more similar to the associative processes young children use? In this study, we compare human and LLMs visual analogy solving. More specifically, we use error analysis to understand *how* LLMs obtain their solutions: is this through abstraction and analogy, association, or perhaps an entirely different process?

Visual analogical reasoning can be assessed in both humans and AI models using the Abstraction Reasoning Corpus (ARC; Chollet, 2019) and the ConceptARC (Moskvichev et al., 2023). The ARC tasks are preferred above other visual reasoning tasks such as the Raven's Progressive Matrices (Raven & Raven, 2003) because these are open-ended rather than multiple-choice and can't easily be solved by chance, the open format also allows better tracing of human and LLM problem representations (Johnson et al., 2021) and the task is designed to assess numerous visual abstractions rather than a limited set of rules (Chollet, 2019; Moskvichev et al., 2023). However, current ARC tasks are too challenging for children as well as LLMs (Moskvichev et al., 2023; Mitchell et al., 2023; Xu et al., 2023). Therefore we created a small set of simplified ARC analogies, inspired by children's visual analogy tasks (e.g., Siegler & Svetina, 2002; Hosenfeld et al., 1997), to gain insights into how LLMs' and children in various phases of analogical reasoning development solve visual analogies.

Young children's visual analogy solving is characterized by associative responses where one of the example images is (partly) duplicated or idiosyncratic responses are given (e.g., choosing a favorite shape and color) (Siegler & Svetina, 2002; Hosenfeld et al., 1997; Stevenson & Hickendorff, 2018). Generally between six and eight years (and younger with training), children start transitioning to successful analogy solving, marked by the appearance of partially correct solutions; here the underlying rule or concept is understood, but one or two aspects are missing (Stevenson & Hickendorff, 2018; Hosenfeld et al., 1997). From 8-years onwards (or with more training) non-analogical responses disappear and (partial) analogical solutions take over (Stevenson & Hickendorff, 2018; Hosenfeld et al., 1997). The transition from associative to analogical reasoning states in visual tasks coincides with what Gentner (1988) refers to as the relational shift. Interestingly, previous work comparing LLM and children analogical reasoning showed that LLMs tend to make similar associative errors as young children(Stevenson et al., 2023).

In the current study we compare LLM and human visual analogy solving, with a focus on exploring whether LLM errors resemble those of young children, adults or represent a less human-like process.

Methods

Data and code are available at `https://github.com/cstevenson-uva/kidsARC`.

KidsARC-Simple and KidsARC-Concept

We created two 8-item sets based on the ARC (Chollet, 2019) and ConceptARC (Moskvichev et al., 2023). The Simple version is geared at younger children (4–8 years) and the Concept version at older children and adults (8+ years).

The original ARC and ConceptARC require few-shot learning with several input-output examples to derive the pat-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* ©2024 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

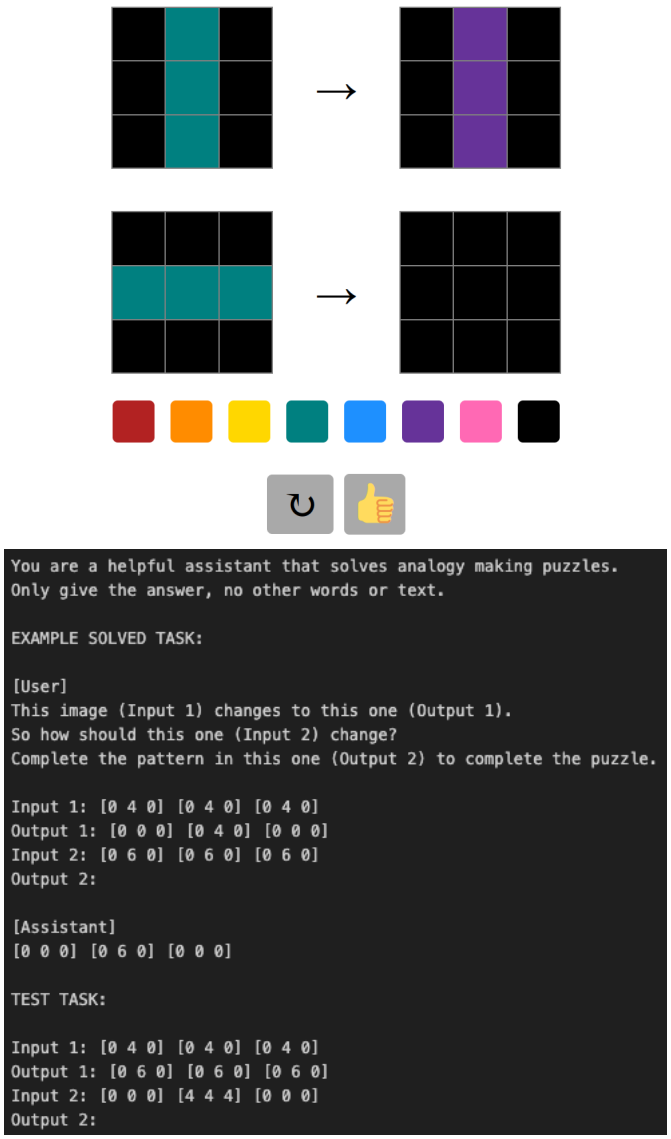

Figure 1: Example item and interface from KidsARC-Simple task (top). Corresponding prompt given to LLMs (bottom), derived from (Moskvichev et al., 2023).

tern and solve the item. We use only one input-output example, thereby mimicking the classical "A is to B as C is to D" analogy set-up, thus requiring one-shot learning (see Figure 1). Having fewer input-output examples poses a challenge in creating unambiguous items, with some items having multiple solutions. However, multiple solutions provide a fruitful ground for testing distinct strategies used by humans and machines, which we discuss later.

KidsARC-Simple items consist of 3x3 grids representing simple concepts such as color and position. KidsARC-Concept consists of 5x5 matrices encoding concepts from the ConceptARC, such as, Inside-Outside and Complete Shape. See Figure 3 for all items.

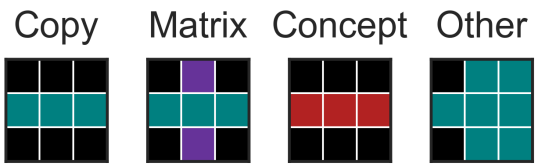

Figure 2: Examples of different error types. See Figure 1 for the full item. A **Copy** error (partially) duplicates an input matrix. A **Matrix** error combines inputs. **Concept** errors apply correct transformations or concepts, but a mistake is made (here the wrong color). **Other** represents idiosyncratic responses.

Human Data Collection

Data collection took place at the NEMO science museum in Amsterdam. The room had several tables with a tablet that ran a cognitive task. Visitors, after signing informed consent, could solve one or more of the five tasks. Two of the tablets were used for this study. Participants received two examples. One to get familiarized with the user interface and one involving a simple analogy. Both examples had written verbal instructions that older children and adults could read and that we read out loud to younger children. All participants saw the items in the same order.

Due to the informal nature and noisy surroundings in the museum, verbal instructions were not always the same across participants. Also, although parents were requested to let their children solve tasks independently and participate themselves in different tasks, they still sometimes (implicitly) helped their children.

We collected data separately for the two item sets. Younger children who wanted to continue after the KidsARC-Simple were allowed to solve the KidsARC-Concept as well. Children solving both tasks were treated as separate participants. Both datasets had a positively skewed age distribution median of 8 (IQR = 7, 10) for KidsARC-Simple (n = 155) and median of 12 (IQR = 9, 35) for the KidsARC-Concept (n = 94). Overall, the youngest participant was 3, and the oldest 76. We binned participants into 5 age bins based on theory expectations: 3-5 (mostly non-analogical), 6-8 (transitioning), 9-11 (mostly analogical), 12+ (successful analogical reasoning) (Stevenson & Hickendorff, 2018).

LLM Data Collection

We used all 60 open-source LLMs available in the Together AI collection. 20 of them failed to be called by the Together AI's API, leaving us with 40 models. Additionally, we used three GPT models from OpenAI: GPT-3 (01.03.2023 snapshot), GPT-4 (13.06.2023 snapshot), and the multimodal GPT-4 Vision (GPT-4V; preview version). Similarly to Moskvichev et al. (2023), we prompted the LLMs by putting the example matrices in text format and coding colors with numbers 0-9. Before presenting each test item, we included the same example item human participants saw, but

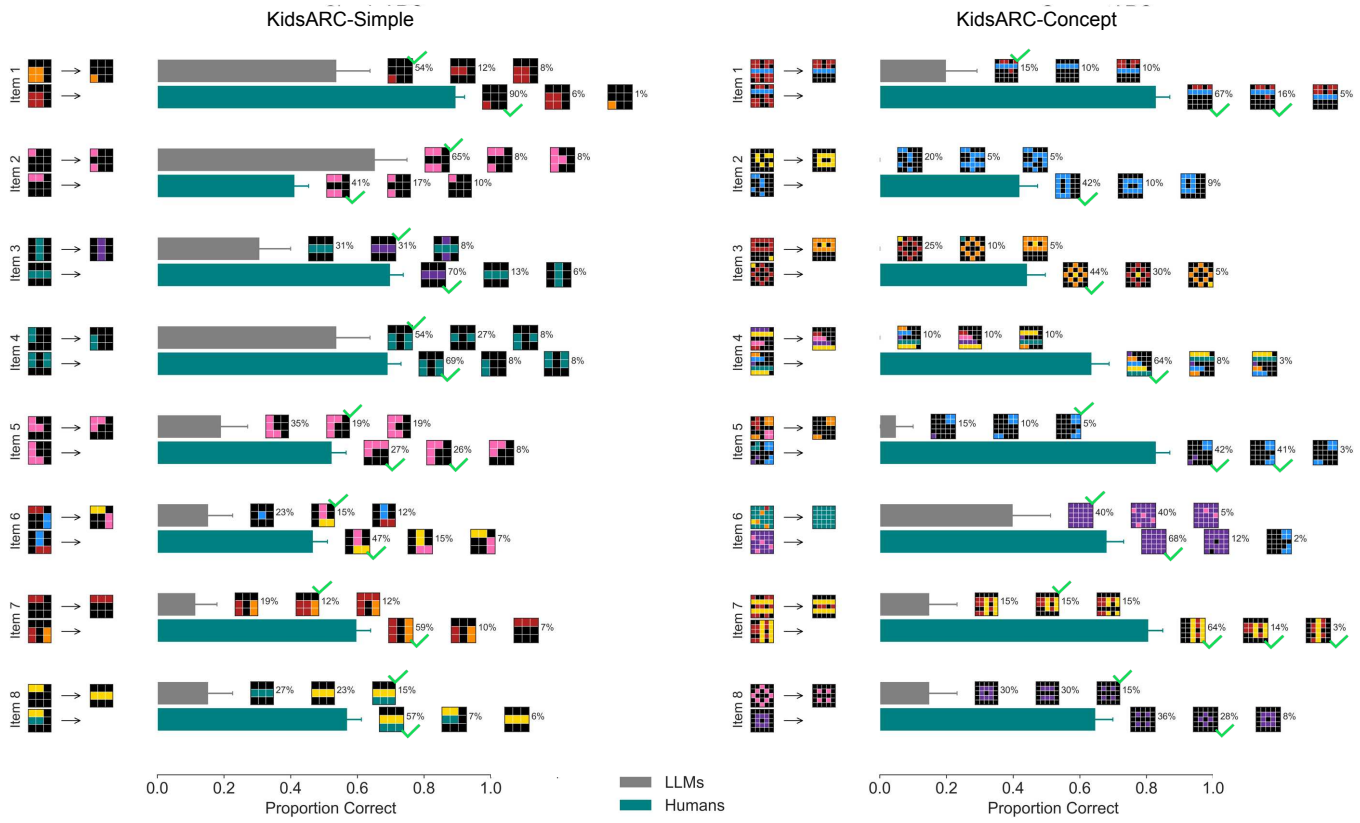

Figure 3: Item-wise comparison of human vs. LLM performance on KidsARC-Simple and KidsARC-Concept. The items are visualized on the left of the performance bars. On the right we display the three most common responses (models and humans) along with percentage occurrence. Green ticks indicate correct responses. Note: items can have more than one correct response.

solved. For GPT's the example item was included in the system prompt, and the test items in the user input. In the case of the LLMs accessed through the Together AI API, the example item was included in the same prompt as the current test item, since the API does not offer system prompt inputs. Similarly to human participants, LLMs were only allowed one attempt at the items and were given no feedback. The temperature was set to 0 for reproducible outputs for all LLMs.

To make use of the image processing capabilities of GPT-4 Vision, we also experimented with presenting the items visually and only requiring text for the output matrix (as with other LLMs) based on the numerical color codes. GPT-4V managed to follow the color coding and output format. However, it performed worse than when using the full text-based prompt that the other models received. Therefore, GPT-4V received the same prompt as other models and the item image was also included to further test its multimodal reasoning capabilities. The temperature setting was not available for GPT-4V at the time of access (20.01.2024).

Exclusion Criteria

Some exclusion criteria were specific to the LLMs, while some were shared across humans and models. We noticed that contrary to GPT models, many LLMs accessed from Together AI produced either incorrectly formatted responses or responses that were completely irrelevant to our items. Some of the common incorrect formatting included either duplicating the same response, providing more than one response, or providing a correctly formatted response followed by text irrelevant to the task. In such cases, we used regular expressions to extract whether a response contained correctly formatted output and if so, we treated the first output as the given response. Otherwise, we classified the responses as a no-response.

After pre-preprocessing LLM responses, we filtered out either human participants or individual LLMs if they had more than two no-responses in a given dataset. In human participants, this simply meant leaving the grid empty, and in the LLMs either a no-response, as defined above, or also an array full of black pixels. This left us with 144 human participants and 26 LLMs in the KidsARC-Simple, and 88 human participants and 20 LLMs in the KidsARC-Concept.

Coding Erroneous Responses

To examine differences in response strategies, we coded all erroneous responses for each item. Based on the literature on children's errors and an exploration of GPT-3's errors, we devised a taxonomy of three kinds of errors: (1) dupli-

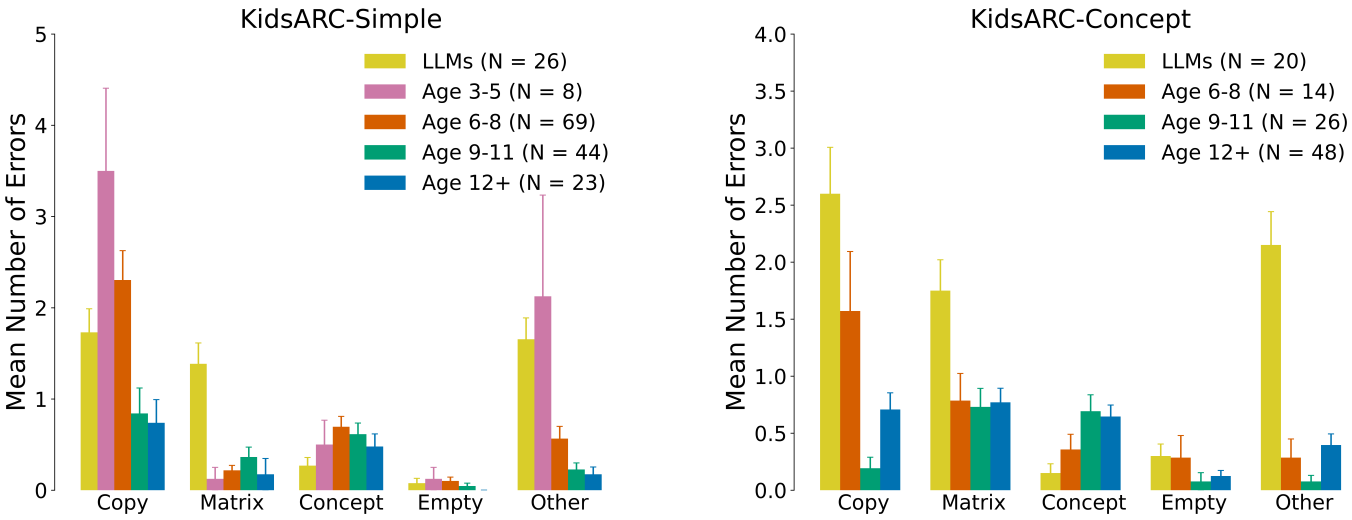

Figure 4: As with young children, copying is a common error in LLMs. Compared to humans, matrix-based errors are more common in LLMs, while concept-based errors are less common. Other errors are common in both LLMs and young children.

cation, (2) concept-based and (3) matrix-based. Duplicating or (mostly[1]) copying one of the analogy elements is the most common error response for children on visual matrix analogies (Stevenson & Hickendorff, 2018; Hosenfeld et al., 1997; Siegler & Svetina, 2002). Concept-based errors are partially correct solutions where the participant clearly understands the concept (e.g., position), but makes a mistake in its application (e.g., moving a pixel too far or forgetting to change the color); such errors are often seen in older children and on difficult items (Stevenson & Hickendorff, 2018; Hosenfeld et al., 1997). Matrix-based errors are a new error type we encountered in this dataset; they are a result of simple matrix combinations of the analogy elements, e.g., the response includes all colored pixels from A, B and C. For example, see the third most common LLM response for item 3 in KidsARC-Simple (Figure 3). The remaining responses were coded as "other" errors. Two independent raters, blinded to whether the response was made by humans or LLMs, coded each response into one of these four categories (initial agreement: 75%), after which all discrepancies were resolved through discussion.

Results

Aggregate Performance Differences

In Figure 3 we summarize the mean performance of all models and all human participants on both the KidsARC-Simple and KidsARC-Concept item sets. On the whole, humans perform better than the models with the exception of one item in the KidsARC-Simple. The difference in performance between the models and humans is even more pronounced when item difficulty increases. For example, on KidsARC-Concept an average LLM does not achieve human performance on any of the items. Also, quite strikingly, there are three items in the KidsARC-Concept that no model solves correctly. There is large variance in how individual LLMs performs, which we discuss in the next section.

Table 1: Comparison of Average Performance (%) between children, adults, and LLMs.

	Age				LLMs
	3-5	6-8	9-11	12+	
KidsARC-Simple	17.2	49.8	73.9	80.4	33.2
KidsARC-Concept	*NA*	56.2	74.0	64.8	11.9

Figure 3 also visualizes the individual items (left) and three most common responses for both the models and humans (next to their respective performance bars). Analyzing individual errors affords insights into the task-solving strategies employed by both humans and models. First, we see that both make the error of literally, or almost literally, copying one of the input matrices. For example, in KidsARC-Simple items 3 and 7 the most common response for the models and the second most common response for the humans was to copy the third input matrix (C-term).

A new strategy we observed is what we call the 'matrix strategy', where the response is a result of some sort of pixel-level merging of the input matrices (see Methods section). This occurred far more often in LLMs. This strategy is well-illustrated by the third most common model responses in items 3 and 7 in the KidsARC-Simple, where the response can be conceptualized as the result of an element-wise union of the input matrices. Formally, if A, B, and C are the input matrices, the response matrix D is given by $D = A \vee B \vee C$, where $\vee$ denotes the element-wise logical OR

[1]Some responses showed slight pixel changes. Nevertheless, exact copies constituted the majority (76%) of the errors classified as copying.

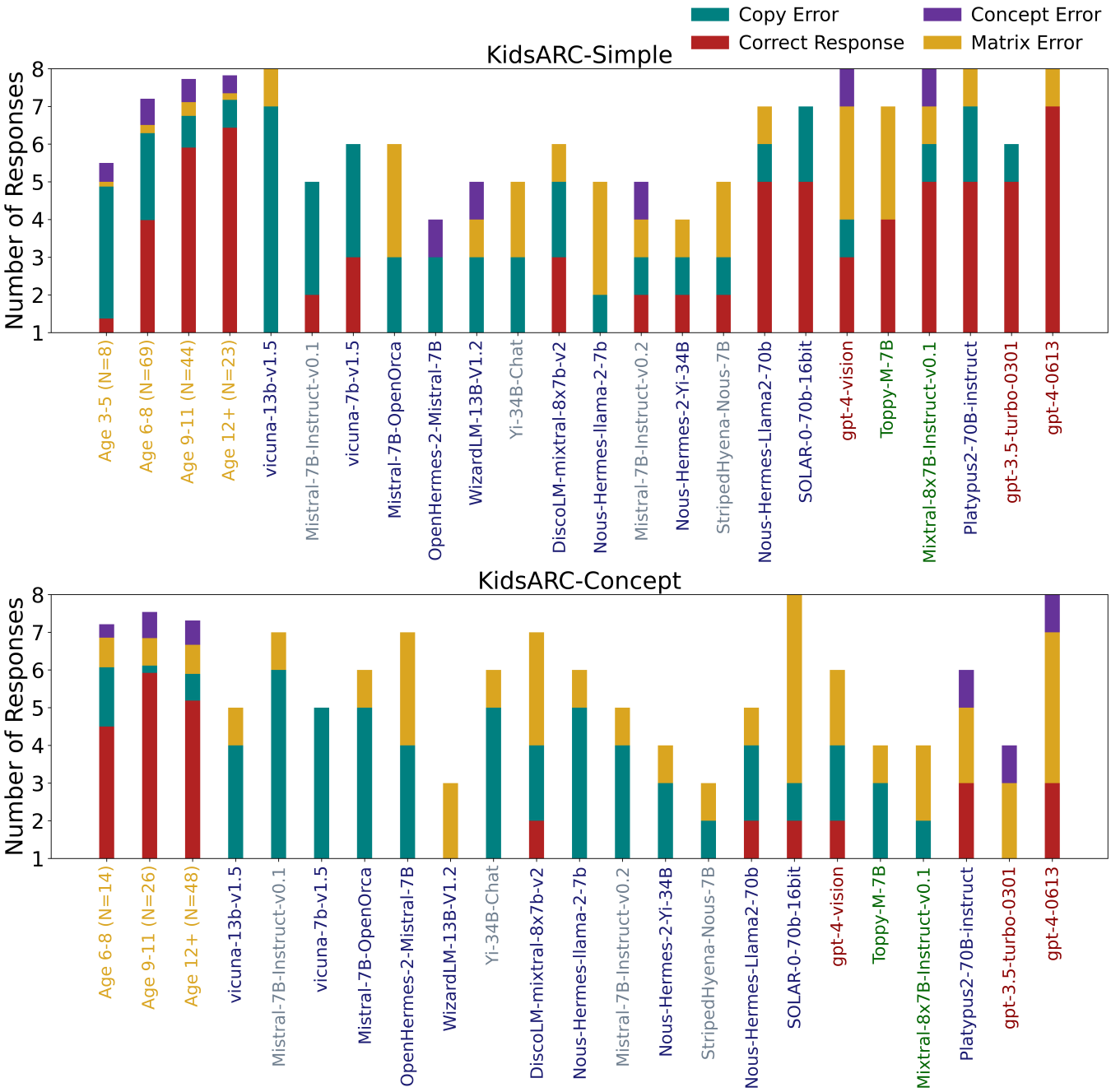

Figure 5: Comparison of accuracy and error types model-by-model and for humans. We plot models with different training regimes - Base Models, Fine-tuned Models, Mixture of Experts, GPTs, as well as Human participants.
Note: To reduce clutter we do not plot 'other', empty, or invalid responses, hence some bars do not sum up to eight.

operation. However, some of the responses that appear to be a result of a matrix strategy, cannot be easily formalized. E.g., in item 6 in the KidsARC-Simple the most common model response can be conceptualized as an XOR-like operation over the three input matrices (black pixels are coded as 0 in the experiment and all other colors are ≥ 1), except the 2nd pixel in the 3rd row (which would be red and not black according to the XOR). Finally, we also observed responses that were a result of direct matrix arithmetic, where LLMs treated the color codes as integers, e.g., in task 3 in KidsARC-Simple - $6(\text{purple}) - 4(\text{teal}) = 2(\text{orange})$.

Errors in LLMs and Humans

In Figure 4 we visualise how often, on average, LLMs and humans across age-groups produce different errors (copy, matrix-based, concept-based, empty responses, or other). We

see that models primarily make copy, matrix, or other errors on both the KidsARC-Simple and KidsARC-Concept items. Similarly, for both younger "pre-analogical" children (3-5 year-olds) and 6-8 year-olds in "transitioning" phases, copying is the most common error and rapidly diminishes in older age-groups. "Other"/idiosyncratic errors are also common for the youngest age-group, but disappear in older groups. Matrix-based errors on the KidsARC-Concept items are more common in LLMs than humans. Concept errors, however, occur very rarely in the models compared to human respondents. The abundance of matrix solution strategies, and lack of concept-based errors shows that LLMs do not generalize abstract concepts as required by the ARC.

Model Spotlight

While there are general trends in LLM responses, there is still much variance in both performance and errors which we explore in Figure 5. Here we plotted the counts of different error types separately for each model. For easy viewing and comparison, we only include models that met the inclusion criteria in the KidsARC-Concept.

We make a couple of preliminary observations. First, there are models that primarily rely on duplication and/or matrix strategies, which are also the models that achieve the lowest performance. The models that perform the best are GPTs (which notably rarely use copying), Mixture of Experts models (although performance dropped in the KidsARC-Concept) and Platypus2-70B-instruct (which was the only model that performed on par with GPT-4 on KidsARC-Concept) . Interestingly, Platypus2-70B-instruct (Lee et al., 2023) is a Llama-2 fine-tuned on a dataset involving logical reasoning tasks which is notable, since the base Llama-2 model performed so poorly that it did not even meet our inclusion criteria. ARC tasks were not part of the dataset.

Models that performed well on either of the tasks, generally had a high parameter count. This is excluding Mixture of Experts models, both having 7B parameters, 10 times less than similarly performing models. On the KidsARC-Simple, however, the smaller 7B parameter Vicuna model (Zheng et al., 2023) performed better than the bigger 13B version.

Discussion

In this paper, we compared how LLMs and children at different stages of analogical reasoning development, perform, and what kind of solution strategies they employed when solving ARC-like items. Our main findings show that LLMs are prone to (partially) copying the input matrices when giving a response - a fallback strategy that young children in pre-analogical and transitioning stages exhibit. What is more, we find that humans and LLMs differ in the types of errors they make. While humans make errors that are conceptually close to the correct solution but might miss a couple of pixels, LLMs often rely on simple combinations of the input matrices. We also found indications of how fine-tuning models on datasets designed to improve reasoning capabilities in LLMs might help in solving visual analogy tasks.

In our investigation, we not only gained insights by looking at error patterns but also recognized the value of including ambiguous items in our dataset. Specifically, two items from the KidsARC-Concept (items 1 and 7, shown in Figure 3) allowed for two valid solutions: applying conceptual knowledge related to spatial relationships (specifically, inside-outside concepts) or through a more straightforward approach of eliminating specific rows or columns. Notably, humans chose the conceptual approach 82% (115/141) of the time, whereas LLMs did so only 14% (1/7) of the time. While the small number of items and low LLM success rate limit strong conclusions, these findings highlight the value of designing ambiguous items for future research into AI and human analogical reasoning capabilities.

Our findings suggest that LLMs, while adept at learning surface statistics, fail to grasp underlying concepts, echoing longstanding critiques of connectionist systems by Fodor & Pylyshyn (1988) and observed in modern AI (Greff et al., 2020). Specifically, neural networks struggle with compositionality and objectness, particularly in tasks like ARC where object-based abstractions are needed for robust generalization (Xu et al., 2024). Similarly, while neural networks are effective at statistical pattern matching, they fail at utilizing abstract structures, unlike humans (Kumar et al., 2023). Our results represent a behavioral example of LLMs' failure to develop symbol-level abstractions, leading to strategies that diverge from those used by humans.

There are a few limitations to consider. First, although we anticipate that measurement errors were mitigated across participants, replicating the experiment under controlled conditions is essential for confirming the reproducibility of our findings with human participants. Second, caution must be exercised when comparing humans to LLMs to avoid pitfalls such as assessing models under conditions dissimilar to those experienced by humans, as highlighted by Ivanova (2023). Despite efforts to standardize item presentation for both groups, inherent differences remain. For instance, presenting the ARC items as matrices to AI models as this and previous studies have done (e.g., Johnson et al., 2021; Moskvichev et al., 2023), may have inadvertently influenced LLMs to adopt simple matrix arithmetic strategies. However, this does not explain why GPT-4-Vision produced many matrix errors despite having access to the tasks in a visual format. Nevertheless, future research should aim to further align human and LLM task presentation.

Analogical reasoning is a cornerstone of human intelligence and creativity (Gentner & Hoyos, 2017). Embedding this capability into AI systems is crucial for achieving generalization capabilities required for robust and trustworthy AI-systems (Mitchell, 2021). In human cognitive development, children exhibit various solution strategy phases in their journey towards proficient analogical reasoning. Our results show that currently LLMs are in the early stages of learning to solve visual analogies and show some non-human deviations.

Acknowledgements

This research was funded by the the Dutch Research Council (NWO) project "Learning to solve analogies: Why do children excel where AI models fail?" with project number 406.22.GO.029 awarded to Claire Stevenson. We thank Arseny Moskvichev and Melanie Mitchell for their valuable feedback on the children's adaptation of the ARC task and Han van der Maas for his insightful contributions to the project. We also thank, in random order, - Ann Sophie Honikel, Jan-Luca Weigand, Milla Rosalina Pihlajamaki, Aaron Benjamin Lob, Leonidas Jung and Konradas Mikalauskas for their attentivenes, persistence, and personal vibrance that implicitly or explicitly added great value to this work.

References

Chollet, F. (2019). On the measure of intelligence. *arXiv:1911.01547*. Retrieved from https://doi.org/10.48550/arXiv.1911.01547

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71. doi: 10.1016/0010-0277(88)90031-5

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, *59*(1), 47–59. doi: 10.1111/j.1467-8624.1988.tb03194.x

Gentner, D., & Hoyos, C. (2017). Analogy and abstraction. *Topics in Cognitive Science*, *9*(3), 672–693. Retrieved from https://onlinelibrary.wiley.com/doi/10.1111/tops.12278 doi: 10.1111/tops.12278

Greff, K., van Steenkiste, S., & Schmidhuber, J. (2020, December). *On the Binding Problem in Artificial Neural Networks.* arXiv. Retrieved 2024-03-18, from http://arxiv.org/abs/2012.05208 (arXiv:2012.05208 [cs])

Hosenfeld, B., van der Maas, H. L., & van den Boom, D. C. (1997). Indicators of discontinuous change in the development of analogical reasoning. *Journal of Experimental Child Psychology*, *64*(3), 367–395.

Hu, X., Storks, S., Lewis, R. L., & Chai, J. (2023). In-context analogical reasoning with pre-trained language models. *arXiv preprint arXiv:2305.17626*. Retrieved from https://arxiv.org/abs/2305.17626

Ivanova, A. A. (2023, December). Running cognitive evaluations on large language models: The do's and the don'ts. *arXiv preprint arXiv:2312.01276*. Retrieved from http://arxiv.org/abs/2312.01276

Johnson, A., Vong, W. K., Lake, B. M., & Gureckis, T. M. (2021). Fast and flexible: Human program induction in abstract reasoning tasks. *arXiv preprint arXiv:2103.05823*. Retrieved from https://arxiv.org/abs/2103.05823

Kumar, S., Dasgupta, I., Daw, N. D., Cohen, J. D., & Griffiths, T. L. (2023, August). Disentangling Abstraction from Statistical Pattern Matching in Human and Machine Learning. *PLOS Computational Biology*, *19*(8), e1011316. Retrieved 2024-04-11, from https://dx.plos.org/10.1371/journal.pcbi.1011316 doi: 10.1371/journal.pcbi.1011316

Lee, A. N., Hunter, C. J., & Ruiz, N. (2023). Platypus: Quick, cheap, and powerful refinement of llms. *arXiv preprint arxiv:2308.07317*. Retrieved from https://arxiv.org/abs/2308.07317

Mitchell, M. (2021). Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, *1505*(1), 79–101. doi: 10.1111/nyas.14619

Mitchell, M., Palmarini, A. B., & Moskvichev, A. (2023). Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks. *arXiv preprint arXiv:2311.09247*. Retrieved from https://arxiv.org/abs/2311.09247

Moskvichev, A., Odouard, V. V., & Mitchell, M. (2023). The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *arXiv:2305.07141*. Retrieved from https://doi.org/10.48550/arXiv.2305.07141

Raven, J., & Raven, J. (2003). Handbook of nonverbal assessment. In (1st. ed., pp. 223–237). Boston, MA: Springer. doi: 10.1007/978-1-4615-0153-4_11

Siegler, R. S., & Svetina, M. (2002). A microgenetic/cross-sectional study of matrix completion: Comparing short-term and long-term change. *Child development*, *73*(3), 793–809.

Stevenson, C. E., & Hickendorff, M. (2018). Learning to solve figural matrix analogies: The paths children take. *Learning and Individual Differences*, *66*, 16–28. Retrieved from https://doi.org/10.1016/j.lindif.2018.04.010 doi: 10.1016/j.lindif.2018.04.010

Stevenson, C. E., ter Veen, M., Choenni, R., van der Maas, H. L. J., & Shutova, E. (2023). Do large language models solve verbal analogies like children do? *arXiv preprint arXiv:2310.20384*. Retrieved from https://arxiv.org/abs/2310.20384

Together.ai. (n.d.). Retrieved 2024-01-29, from https://docs.together.ai/reference/chat

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*, 1526–1541. Retrieved from https://www.nature.com/articles/s41562-023-01659-w

Xu, Y., Li, W., Vaezipoor, P., Sanner, S., & Khalil, E. B. (2023). Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *arXiv preprint arXiv:2305.18354*. Retrieved from https://arxiv.org/abs/2305.18354

Xu, Y., Li, W., Vaezipoor, P., Sanner, S., & Khalil, E. B. (2024, February). *LLMs and the Abstraction and Reasoning Corpus: Successes, Failures, and the Importance of Object-based Representations.* arXiv. Retrieved 2024-05-07, from http://arxiv.org/abs/2305.18354 (arXiv:2305.18354 [cs])

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., … Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv:2306.05685*. Retrieved from `http://arxiv.org/abs/2306.05685`