Are LLMs tools to understand human neurocognition during abstract reasoning?

Christopher Pinier (c.pinier@uva.nl)
Claire E. Stevenson (c.e.stevenson@uva.nl)
Michael D. Nunez (m.d.nunez@uva.nl)
Psychological Methods, University of Amsterdam
Nieuwe Achtergracht 129-B
1018 WS Amsterdam, The Netherlands

Abstract

Abstract reasoning, a key component of human intelligence, seems to have recently emerged in large language models (LLMs). If so, LLMs could help us provide a mechanistic explanation for the brain processes behind the abstract reasoning abilities of humans. In this study, we compared the performance of multiple LLMs to human performance in a visual abstract reasoning task. We found that while most LLMs cannot perform this task as well as human participants, some LLMs are competent enough for use as potential descriptive models. We propose that the best-performing LLMs can be used as models to understand human performance, response times, and the timing of Event-Related Potentials (ERPs) as recorded by electroencephalography (EEG) during the task. We show initial behavioral and ERP results, and present our plan to compare LLM embeddings and surprisal measures to cortical activity patterns. This is the first step in a larger project to create neurally-informed artificial networks as tools to understand human neurocognition.

Keywords: AI; Abstract Reasoning; Artificial Neural Networks; Deep Learning; Large Language Models (LLMs); EEG; Event-Related Potentials (ERPs)

Introduction

Recent advancements in Artificial Intelligence (AI), driven by the development of massive, closed-source models, have sparked global excitement about the field as well as an ongoing debate about the genuine reasoning capabilities of these systems. On the one hand, some research indicates emerging analogical and abstract reasoning abilities in LLMs (Webb, Holyoak, & Lu, 2023). On the other hand, these models could fool us by merely exploiting statistical patterns that are near to fully imperceptible to humans (Kumar, Dasgupta, Daw, Cohen, & Griffiths, 2023), or even worse, by regurgitating examples that were present in a contaminated training dataset (Wu et al., 2023).

Until the advent of LLMs, the abilities of AI relative to humans were mainly investigated within the field of vision neuroscience (Bankson et al., 2018; Cichy et al., 2016; Eickenberg et al., 2017), with representational similarity analysis (RSA) being a pivotal tool in revealing correspondences between the human brain and deep neural networks. This technique involves comparing the similarity between activation patterns across different layers of neural networks and brain activity across various regions to understand how each processes information. For instance, studies have shown that certain layers in convolutional neural networks engaged in visual tasks exhibit activation patterns that closely resembled those in the human visual cortex, particularly in tasks involving object recognition and categorization (Güçlü & Van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Kubilius et al., 2019; Xu & Vaziri-Pashkam, 2021; Zeman et al., 2021). While such studies have led to important advances, they have been restricted to perceptual processes and their conclusions have been moderated by other work highlighting important differences in the way these models process visual information (e.g., sensitivity to adversarial attacks, texture bias, etc.; for a review, see paper by Bowers et al. (2023)).

We plan to extend this work towards understanding how the brain supports higher-level cognition, by studying how activation patterns in layers of LLMs correlate with neurocognition during reasoning, measured using human behavior and EEG. We use a simple task where the goal is to complete the pattern in a series of icons.

The current study focuses on investigating which LLMs are candidate models of the human reasoning process by comparing LLM performance to that of a human sample. We have identified models with human-like performance on our task, and in the near future, we will examine whether these models indeed predict the timing and magnitude of Event-Related Potentials (ERPs) during abstract reasoning.

Materials and methods

Participants A total of 60 participants were recruited from the online platform Prolific, comprising an equal gender distribution (50% female) with an average age of 37.8 years (SD = 12.2).

Large Language Models Responses were collected from multiple open-source LLMs available via *together.ai*'s public API (available at `https://www.together.ai/`). Responses were also collected from ChatGPT (OpenAI) and Claude (Anthropic), through their respective online chat interface.

Task design On each trial, participants are presented with a series of icons arranged in a way that follows a specific pattern (e.g., 'ABABABAB'), while LLMs are presented with corresponding word descriptions. The goal is to predict the next icon (or word) in the series, selecting from four multiple-choice options. In the task given to participants, all icons and four response options are presented simultaneously until a response

is made (mouse click) or the maximum response window (15 seconds) has been reached (see Figure 1). In the pilot EEG experiment, the task is similar but each icon is first briefly presented (600 milliseconds) to record icon encoding before the whole sequence along with the four options are shown at once.

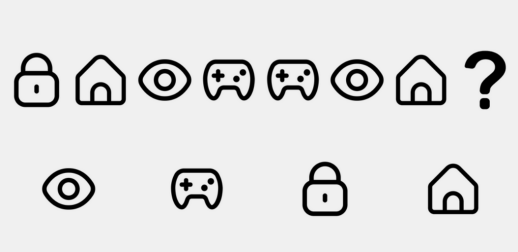

Figure 1: Experimental sequence example. Top row: the series to be completed; bottom row: four response options; correct answer: lock icon (third position).

EEG Apparatus Pilot EEG data were collected using a 64-channel headcap (10-20 layout) from BioSemi, connected to an EEG amplifier system with a sampling rate of 2048 Hz.

Data Analysis

LLMs We compared the models' choice accuracy to that of human participants on the whole and for each abstract pattern.

EEG One participant's EEG data during a similar version of the visual abstract reasoning task written in PsychoPy. This data was collected to pilot the overall study. We cleaned the data using a [1, 10] bandpass filter and an automatic cleaning procedure relying on Independent Component Analysis. Visual Event-Related Potentials (ERPs) were calculated for each task phase (encoding phases vs. reasoning phase). We observed N200s (negative peaks around 200 ms in posterior electrodes, see Nunez, Gosai, Vandekerckhove, and Srinivasan (2019)), to each visual onset.

Results

Human Performance On average, human participants reached an accuracy of 85.60% across all patterns. Accuracy tends to be high ($> 90\%$) on simpler patterns like 'AAAAAAA' or 'AAABAAAB' but decreases on more complex ones, such as 'ABCDEEDC' (dropping to about 53.69%).

LLMs Performance GPT-4 and Claude 3 Sonnet exhibit the highest average accuracy, close to human levels at approximately 81.58% and 83.33% respectively. Other models, like the Nous-Hermes-2-Yi-34B and various Qwen models, perform significantly lower, with scores ranging from around 44.74% to 56.14%.

Pattern-Specific Performance Both humans and LLMs show variability in performance across different patterns. Notably, all models and humans scored highly on the simplest

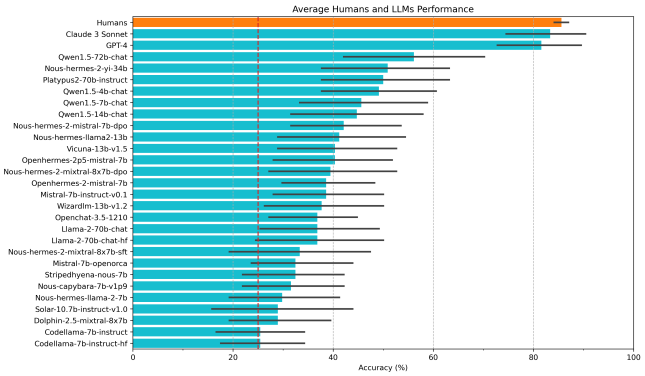

Figure 2: Average performance of human participants and LLMs across all pattern types. Red dotted line: chance-level performance.

pattern ('AAAAAAA'), often reaching perfect scores. However, on complex patterns like 'ABCDEEDC', performance notably decreases for both humans and LLMs. This drop, however, is more dramatic for LLMs. For example, between 'AAAAAAA' and 'ABCDEEDC' sequences humans dropped from 97.54% to 53.69% models (excluding those who performed on average below chance-level) dropped from 70.24% to 14.29%.

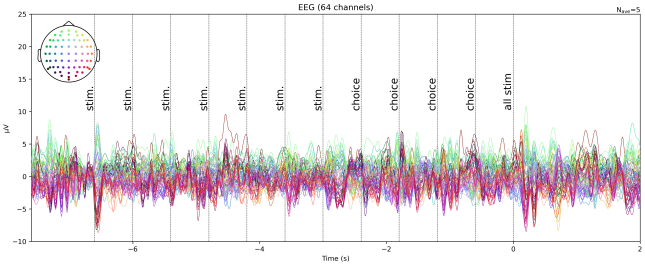

Figure 3: EEG activity timeseries showing ERPs in each encoding phase and the final reasoning phase in the pilot participant.

Discussion

Our results show that most LLMs perform below human level accuracy. However, the two closed-source models we tested performed similarly to humans. Yet because these models are closed-source we cannot use activation patterns in layers of these LLMs to predict the ERPs. However, it is possible to compute surprisal based on an LLM's log probabilities for each pattern completion option, which could be used to predict ERPs (see Figure 3) with a slightly altered task design. Once we have a model that can predict ERPs during the reasoning phase, we plan to work towards creating a neurally-informed artificial networks as tools to understand human neurocognition.

References

Bankson, B. B., Hebart, M. N., Groen, I. I. A., & Baker, C. I. (2018). The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *NeuroImage*, *178*, 172–182. doi: 10.1016/j.neuroimage.2018.05.037

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... others (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755. doi: 10.1038/srep27755

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194. doi: 10.1016/j.neuroimage.2016.10.001

Güçlü, U., & Van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., ... DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. , *32*.

Kumar, S., Dasgupta, I., Daw, N. D., Cohen, J. D., & Griffiths, T. L. (2023). Disentangling abstraction from statistical pattern matching in human and machine learning. *PLoS computational biology*, *19*(8), e1011316.

Nunez, M. D., Gosai, A., Vandekerckhove, J., & Srinivasan, R. (2019). The latency of a visual evoked potential tracks the onset of decision making. *Neuroimage*, *197*, 93–108.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*, 1526–1541.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., ... Kim, Y. (2023). Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*.

Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. , *12*(1), 2065. (Publisher: Nature Publishing Group) doi: 10.1038/s41467-021-22244-7

Zeman, A. A., Ritchie, J. B., Bracci, S., & Op de Beeck, H. (2021). Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. , *10*(1), 2453. (Publisher: Nature Publishing Group) doi: 10.1038/s41598-020-59175-0