Moderate evidence for large language models reflecting 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

Large language models (LLMs) have shown alignment with human brain activity during language tasks, but it remains unclear whether this correspondence extends to higher-order cognition such as abstract reasoning. In this study, we compared human EEG responses—specifically fixation-related potentials (FRPs) time-locked to gaze fixations onset —to the internal activations of eight open-source LLMs performing a visual abstract reasoning task. Intermediate LLM layers showed clear differentiation across reasoning pattern types, suggesting potential specialization. While the best-performing models reached human-level accuracy, they did not consistently align with human behavioral patterns. Representational similarity analysis revealed only moderate correlations between model activations and FRP data. This may reflect a lack of neural alignment in LLMs and/or that there is only some relevant cognitive signal in the FRPs. These findings highlight both the promise and limitations of using LLMs as models of human abstract reasoning.

Keywords: AI; Abstract Reasoning; Artificial Neural Networks;Large Language Models (LLMs); EEG; Fixation-Related Potentials (FRPs)

Introduction

The past three decades have seen artificial intelligence systems surpassing human abilities in specialized tasks, including board games, video games, and complex biological problems like protein folding (Silver et al., 2016; Mnih et al., 2015; Abramson et al., 2024). Early breakthroughs with convolutional neural networks, notably AlexNet, demonstrated remarkable similarities between neural network layers and human visual processing, specifically in terms of hierarchical feature extraction from low-level edge detection to higher-level shape recognition (Krizhevsky et al., 2012; Cichy et al., 2016; Yamins et al., 2014).

The emergence of transformer-based Large Language Models (LLMs), such as ChatGPT, marked a significant shift, showcasing impressive capabilities in human-like language comprehension and reasoning tasks. Recent studies revealed substantial correspondence between internal representations of these models and neural activity recorded during linguistic and semantic tasks in humans (Schrimpf et al., 2021; Caucheteux et al., 2022). Larger and more sophisticated models, in particular, showed stronger alignment with neural signatures, including the prediction of neural responses such as the N400 event-related potential (Huber et al., 2024). Such alignment suggests that brain-like processing could be an emergent property of optimizing language understanding.

However, it remains unclear whether this representational alignment also extends to higher-order cognitive functions, such as abstract reasoning. Here, we directly address this question by comparing eye-fixation related neural representations recorded with EEG to internal activations within several LLMs solving analogous abstract reasoning problems.

Materials and methods

Task

Participants were seated in front of a computer screen while EEG signals were recorded (64-channel BioSemi; 2048 Hz) simultaneously with eye tracking (EyeLink 1000 Plus; 2000 Hz). On each trial, participants performed an abstract reasoning problem in which they viewed a sequence of icons arranged according to an implicit logical rule (a “pattern”, such as ABBAABBA). The final icon was masked by a question mark and the goal was to select, using the keyboard, the correct continuation out of four possible options displayed below.

The experiment comprised 400 unique trials divided into 5 sessions of 80 trials each, with each session containing 10 trials per pattern type.

Large Language Models

We evaluated eight open-source, instruction-tuned LLMs of varying sizes: Llama (3.2-3B, 3.3-70B), Gemma (2-2b, 2-9b, 2-27b), Qwen2.5-72B, DeepSeek-R1-Distill-Llama-70B, and Microsoft Phi-4. These models were downloaded and run locally using the Hugging Face library and activations from each of their hidden layers were exctracted. Each LLM was queried one prompt at a time using a one-shot prompting strategy on a text-based version of the task, with each icon replaced by a one-word label. Additional instructions were provided to try and ensure adherence to a specific response format and a clear understanding of the task.

Data Analysis

EEG representations FRPs to each icon of a trial’s sequence were extracted (-100 to 600 ms relative to fixation onset) and averaged together across occipital electrodes and

across fixations per trial, thus producing a trial-level representation of EEG activity related to the neurocognitive processing of icons during abstract reasoning. It was thought that these FRPs represented snapshots into the cognitive processing of pattern completion. We also leveraged the fact that the signal-to-noise ratio of EEG improves by averaging over multiple segments time-locked to fixations within one experimental trial.

LLM-layer representations We extracted a subset of each layer's activations to isolate those related specifically to the tokens of a given sequence, producing a trial-level representation of a layer's activity to that sequence.

Representational Similarity Analysis (RSA) Representational Dissimilarity Matrices (RDMs) were generated from the FRP time-samples and LLM layers' activations data at the pattern-level, that is, averaged across sequence representations pertaining to the same pattern type.

The layer RDMs of each model were compared to an "ideal" reference RDM, consisting of 0's for diagonal elements and 1's for off-diagonal elements. The layer with the highest similarity to this reference RDM was then used in the comparison with human RDMs.

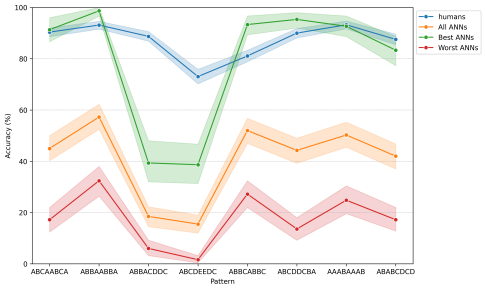

Figure 1: Accuracy By Pattern Type. Best LLMs performed above 60% overall, Worst LLMs below 60%.

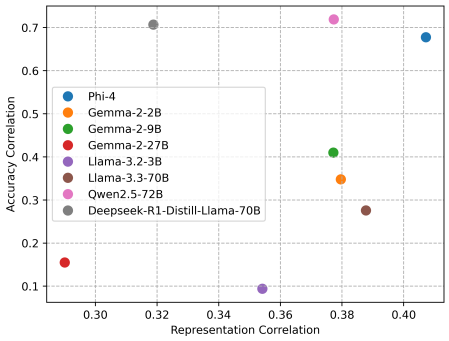

Figure 2: Relationship between FRP (x-axis) and accuracy (y-axis) alignment with Humans across LLMs.

Results

Behavioral Alignment In terms of performance on the task, Llama-3.3-70B achieved the highest accuracy (81.75%), closely followed by Qwen2.5-72B (80.50%) and DeepSeek-R1-Distill-Llama-70B (75.00%). All other models fell below 40.00%.

However, high accuracy did not necessarily equate to human-like behavior. While both Qwen2.5-72B and DeepSeek-R1-Distill-Llama-70B showed a relatively high correlation to the accuracy by pattern of the human group (0.72 and 0.71, respectively), Llama-3.3-70B, the top performer, only showed a weak correlation (0.27). The latter was even surpassed by Phi-4, a low-performing model displaying a 0.67 correlation with the human group.

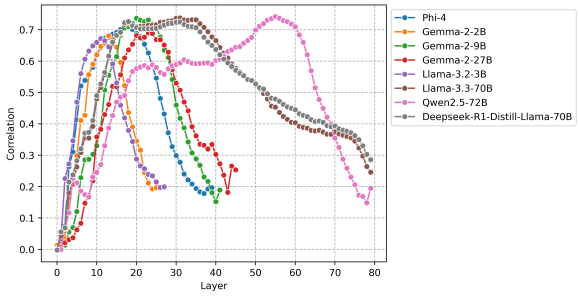

Figure 3: Layer-wise similarity between LLM representations and an idealized reference RDM.

Representational Alignment The RSA comparing activations of the best layer to human FRP data yielded moderate correlations (about 0.3 to 0.4) for all models. Models with higher behavioral alignment did not show stronger similarity with neural data. Qwen2.5-72B and DeepSeek-R1-Distill-Llama-70B, which demonstrated behavioral performances similar to that of humans , did not differentiate themselves from the other models here, with correlations of 0.39 and 0.32, respectively. Interestingly, Phi-4, with an overall accuracy of 32.00% on the task, ranks highest in representational similarity with neural data, again surpassing better performing LLMs.

Discussion

Our results showed moderate representational similarity between human FRP data and LLM activations, but without a clear advantage for higher-performing or behaviorally aligned models. Interestingly, strong task performance did not necessarily predict human-like patterns of accuracy. This relatively inconclusive representational alignment might be due to limitations in the FRPs, which might have been too noisy or insufficiently sensitive to capture the cognitive processes underlying abstract reasoning. In contrast, intermediate LLM layers consistently displayed clear differentiation of abstract pattern types, regardless of task performance. Future work should therefore investigate the computational roles of these specialized layers to clarify their function and relevance in reasoning tasks.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., ... Jumper, J. M. (2024, June). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, *630*(8016), 493–500. (Nature Publishing Group) doi: 10.1038/s41586-024-07487-w

Caucheteux, C., Gramfort, A., & King, J.-R. (2022, September). Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, *12*(1), 16327. (Nature Publishing Group) doi: 10.1038/s41598-022-20460-9

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

Huber, E., Sauppe, S., Isasi-Isasmendi, A., Bornkessel-Schlesewsky, I., Merlo, P., & Bickel, B. (2024, April). Surprisal From Language Models Can Predict ERPs in Processing Predicate-Argument Structures Only if Enriched by an Agent Preference Principle. *Neurobiology of Language*, *5*(1), 167–200. doi: 10.1162/nol_a_00121

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems* (Vol. 25). Curran Associates, Inc.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015, February). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. (Nature Publishing Group) doi: 10.1038/nature14236

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021, November). The neural architecture of language: integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(45), e2105646118. doi: 10.1073/pnas.2105646118

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016, January). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. (Nature Publishing Group) doi: 10.1038/nature16961

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014, June). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. (Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1403112111