UNIVERSITY OF AMSTERDAM
Faculty of Social and Behavioural Sciences

Are Large Language Models Adequate Tools to Understand Human Abstract Reasoning?

Christopher Pinier, Claire E. Stevenson, Michael D. Nunez

Abstract reasoning: involves identifying patterns, deriving general rules from specific examples, and applying flexible thinking to develop solutions in unfamiliar scenarios

Some research indicates emerging analogical and **abstract reasoning** abilities in **Large Language Models** (LLMs) [1]

However these models could be:

- **exploiting statistical patterns** that are near to fully imperceptible to humans [2]
- **regurgitating examples** present in a contaminated training dataset [3]

METHODS

Task

- Series of **icons** arranged in **specific patterns** (e.g., "ABABABAB", "AAABAAAB")
- Goal: **predict the last icon** in the sequence

Participants

- 25 adults
- **EEG** (64 channels) + **Eye-Tracking**

LLMs

- 8 open-source LLMs tested on **text-based version** of the same task using **one-shot prompts**
- Activations extracted from **every hidden layer**

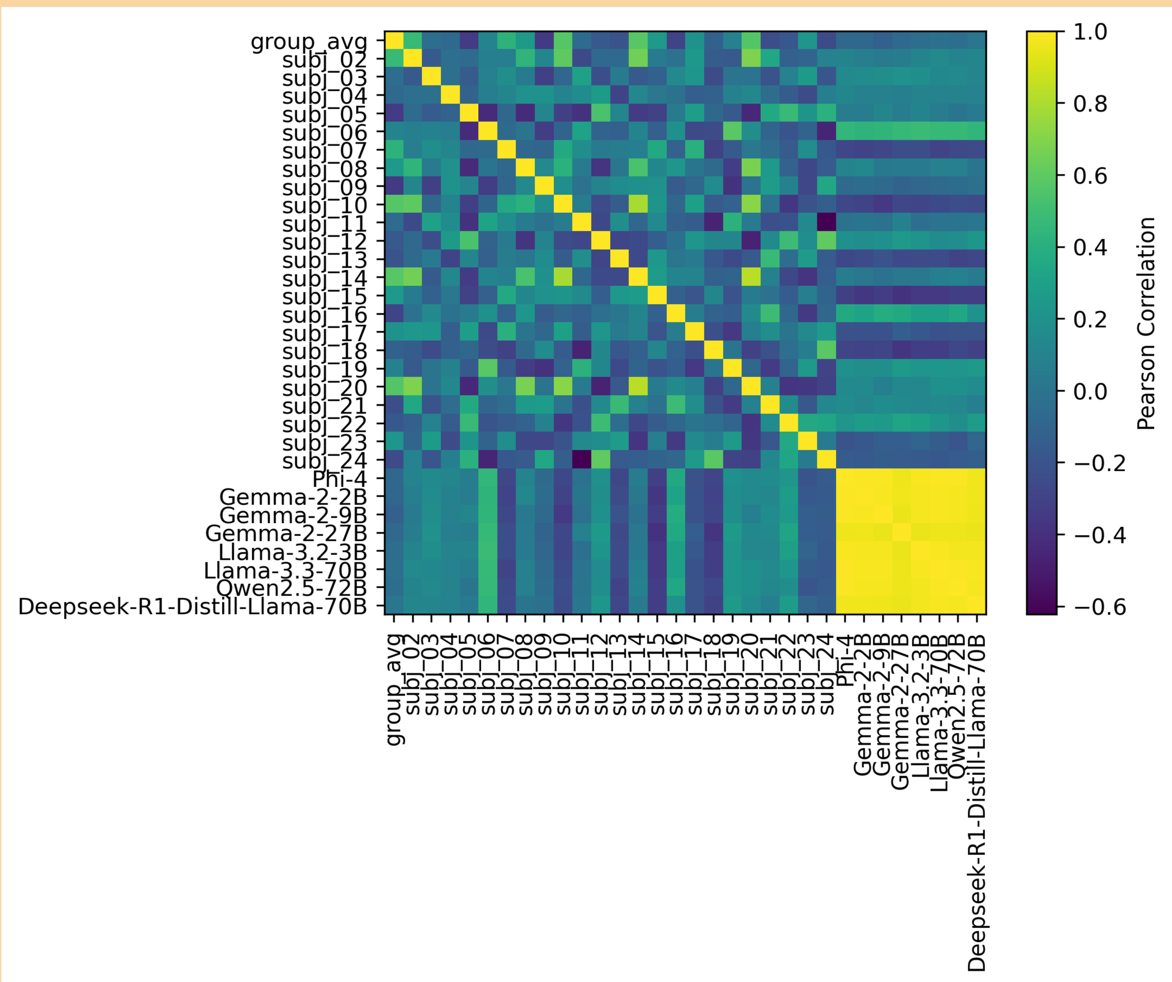
Analysis

- **Fixation-Related Potentials (FRPs)** extracted from frontal electrodes during gaze fixations on each series' icons
  - Chosen for their **ecological validity**: reflect real-time, self-paced processing
  - Compared against traditional **ERPs** time-locked to **response onset**
- **Representational Similarity Analysis (RSA)** used to compare:
  - **FRPs** vs. **LLM activations**
  - **ERPs** vs. **LLM activations**
  - **Human accuracy** vs. **LLM accuracy**
- Similarity quantified via **correlation of RDMs**, with **permutation tests** for statistical significance

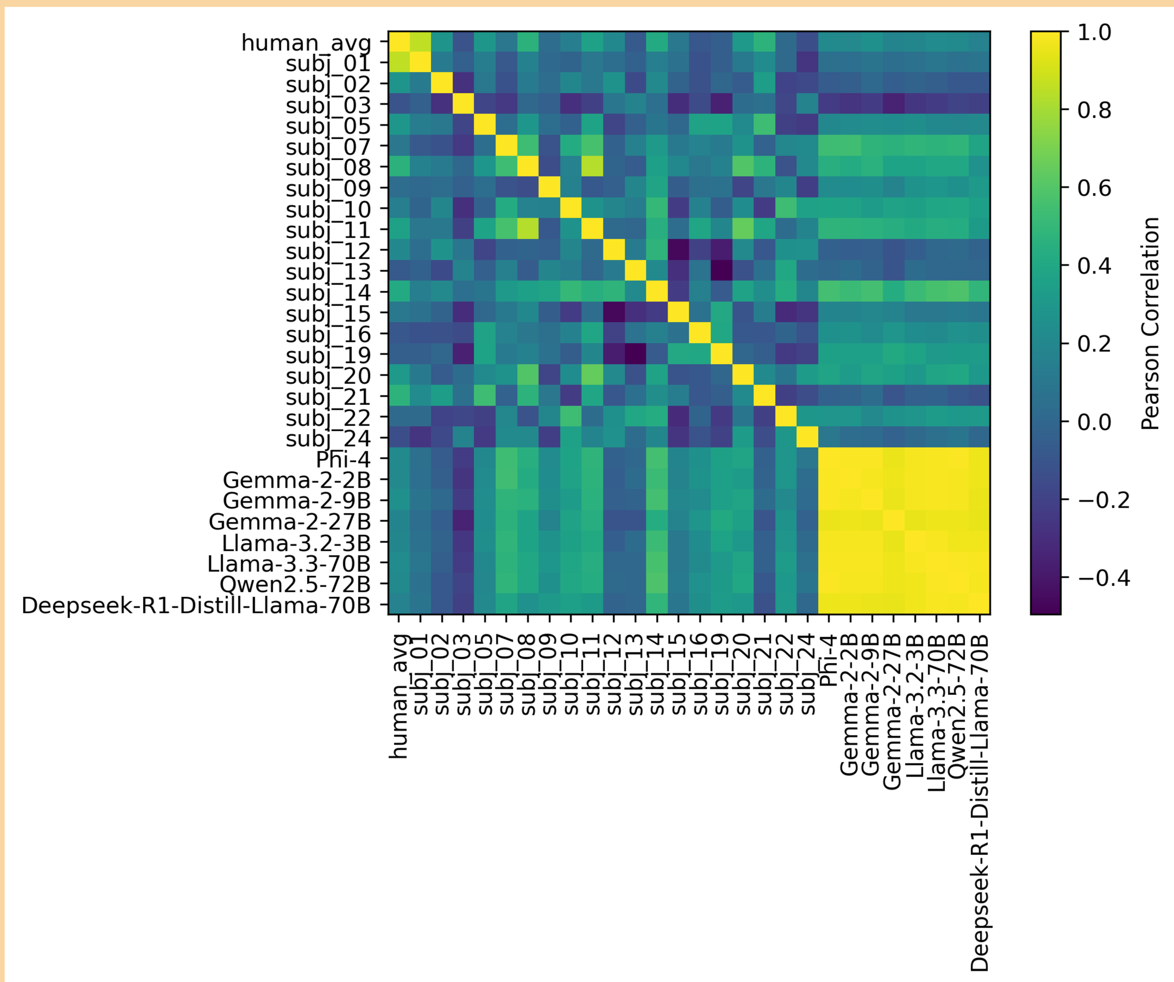
Some LLMs show **human-like reasoning behavior** and organize their **internal states by abstract structure** — potentially reflected in Fixation-Related Potentials

Follow the project on github

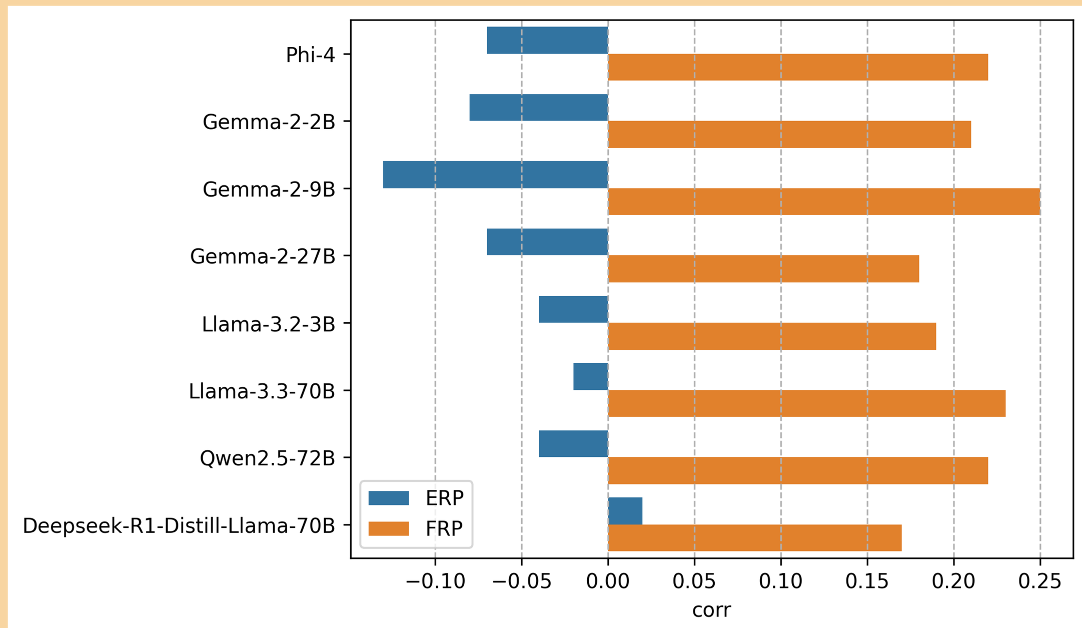
Representational Similarity Analysis



Similarity Matrix from ERP data



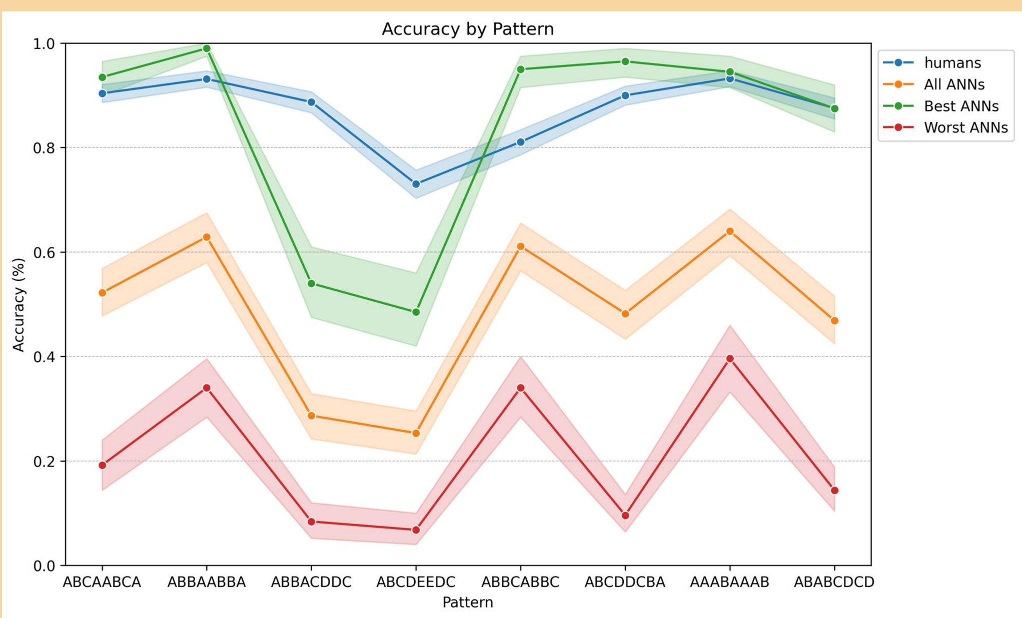
Similarity Matrix from FRP data



EEG–LLM Representational Similarity (ERP vs. FRP)

- **LLMs cluster tightly** among themselves, but show **limited alignment with human neural data**
- Human **cortical representations** are much **more heterogenous**
- Heatmaps show **higher within-human consistency** for FRPs vs. ERPs

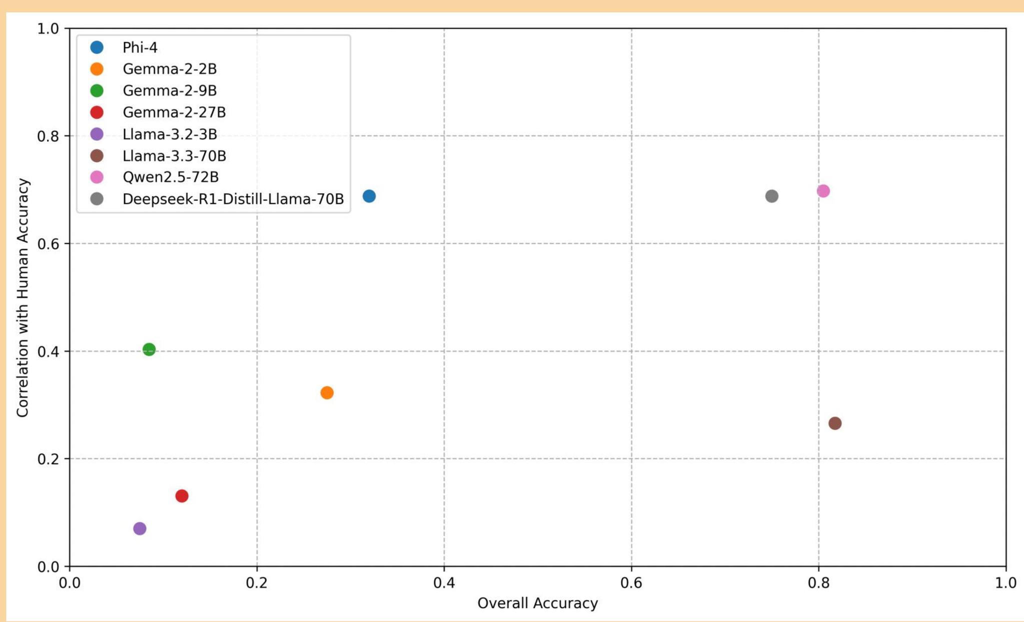
- **Fixation-Related Potentials (FRPs)** showed **modestly higher alignment** with LLM activations compared to **ERPs**
- **No correlations reached significance** (all p > .05 on permutation test), but:
  - FRP data showed **consistent positive correlations** across models
    - r = .17 to .25 (M = - 0.05 , SD = 0.04)
  - ERP data showed **near-zero or negative** values, indicating **weaker representational overlap**
    - r = –.13 to +.02 (M = 0.21 , SD = 0.03)



Accuracy by Pattern Type

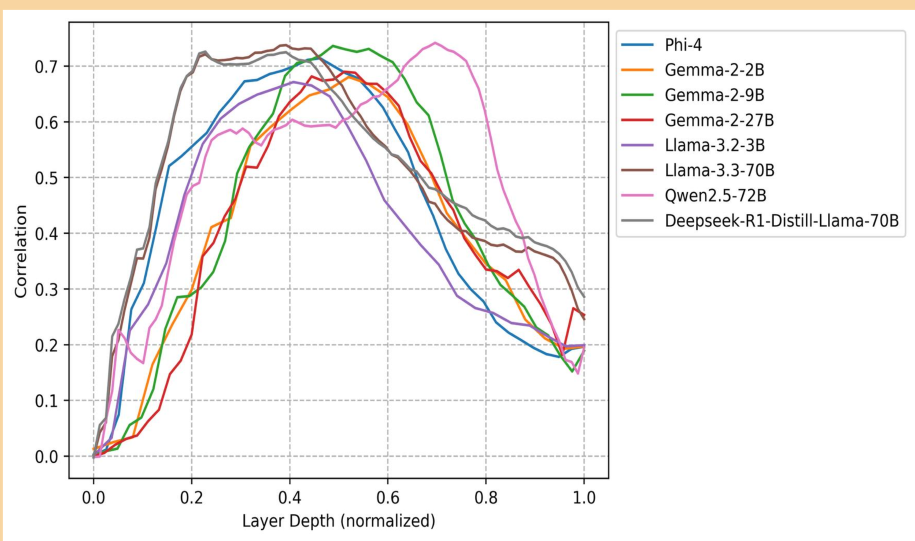
Green: Best LLMs (*accuracy > 0.6*): Qwen2.5-72B, DeepSeek-70B, Llama-3.3-70B
Red: Worst LLMs (*accuracy < 0.6*): all remaining models

- **Humans** consistently **outperform** all model groups, but show pattern-specific difficulty (e.g., *ABCDEEDC*)
- **Best LLMs** show a **human-like accuracy profile**, with better alignment on dips and peaks
- **Worst LLMs** perform poorly overall and exhibit **inconsistent profiles**, suggesting weaker pattern sensitivity

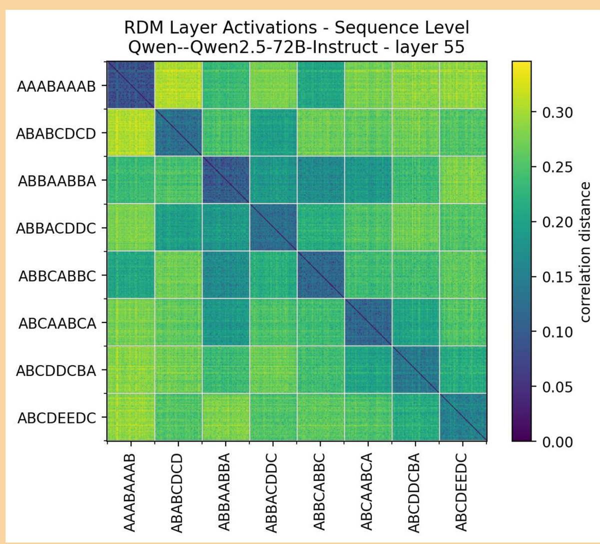


Task Accuracy vs. Human-Likeness

- Most LLMs tested **do not perform well** (clustered in **lower-left quadrant**)
- **Ideal candidates** for modelling human cognition lie in the **top-right**:
  - High task performance and high similarity to human response patterns
- **Qwen2.5-72B** and **Deepseek-R1-Distill-Llama-70B** are promising models, combining:
  - **High accuracy** (≥ 75%)
  - **Strong correlation** with human response patterns (r > .70)
- **Llama-3.3-70B**, despite being most accurate (82%), **aligns poorly** with human response structure (r = .27)
- **Phi-4** shows the **opposite profile**: low accuracy, but surprisingly human-like perfomance (r = .67)



Correlation with an **idealized reference RDM** that encodes **pattern identity**



Example **RDM of LLM layer** with **highest similarity** to reference RDM

Intermediate layers in LLMs appear to **encode task-relevant abstract structure** most strongly
=> indicates that **pattern membership** becomes an **explicit organizing principle** in these layers

Clear **block structure** show that internal representations cluster by **abstract pattern type**

CONCLUSION

- All tested LLMs develop **pattern-sensitive internal representations**, and a subset of them show **human-like response structure**
- **FRPs** seem to provide **a more ecologically valid neural correlate than ERPs**, yielding **more stable alignment** with LLM representations

LIMITATIONS

- **No statistically significant** brain–model alignment found (p > .05), possibly due to limited statistical power or signal resolution
- Analysis **did not yet leverage known neural markers** (alpha-beta phase amplitude
- coupling, N400, etc.)

FUTURE DIRECTIONS

- **Mechanistic Interpretability** on top-performing open-source LLMs to identify components (e.g., attention heads, subnetworks) **functionally relevant to abstract reasoning**
- **Eye movement strategy** analysis (e.g., gaze transitions, heatmaps) compared to LLM **attention weights** for insight into reasoning paths

REFERENCES
[1] Webb (2023), Nature Human Behaviour, 7.
[2] Kumar (2023), PLoS computational biology, 19.
[3] Wu (2023), arXiv preprint, arXiv:2307.02477.