LARGE LANGUAGE MODELS SHOW SIGNS OF ALIGNMENT WITH HUMAN NEUROCOGNITION DURING ABSTRACT REASONING

Christopher Pinier[*], **Sonia Acuña Vargas, Mariia Steeghs-Turchina,**
Dora Matzke, Claire E. Stevenson[†] **& Michael D. Nunez**[†]
Psychological Methods
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

This study investigates whether large language models (LLMs) mirror human neurocognition during abstract reasoning. We compared the performance and neural representations of human participants with those of eight open-source LLMs on an abstract-pattern-completion task. We leveraged pattern type differences in task performance and in fixation-related potentials (FRPs) as recorded by electroencephalography (EEG) during the task. Our findings indicate that only the largest tested LLMs ($\sim$70 billion parameters) achieve human-comparable accuracy, with Qwen-2.5-72B and DeepSeek-R1-70B also showing similarities with the human pattern-specific difficulty profile. Critically, every LLM tested forms representations that distinctly cluster the abstract pattern categories within their intermediate layers, although the strength of this clustering scales with their performance on the task. Moderate positive correlations were observed between the representational geometries of task-optimal LLM layers and human frontal FRPs. These results consistently diverged from comparisons with other EEG measures (response-locked ERPs and resting EEG), suggesting a potential shared representational space for abstract patterns. This indicates that LLMs might mirror human brain mechanisms in abstract reasoning, offering preliminary evidence of shared principles between biological and artificial intelligence.

1 INTRODUCTION

The prospect of developing an intelligent system akin to human intelligence has long captivated the public imagination. Once relegated to science fiction, it became tangible with early computer systems and later milestones such as computers mastering Chess and Go (Silver et al., 2016; 2017; 2018) and the release of ChatGPT. Large language models (LLMs) now approach human-level competence across diverse reasoning tasks, yet whether these models reason in a genuinely human-like manner remains an open question. Abstract reasoning, a fundamental ability to extract patterns, rules, and relationships from limited information and apply them in new contexts, is widely regarded as a cornerstone of human cognition and provides a stringent test for this question. Early AI pioneers already framed intelligence as abstraction and rule manipulation (Newell, 1955; Newell & Simon, 1956), and recent LLMs achieve strong performance on abstract-reasoning tasks (Bubeck et al., 2023; Musker et al., 2025; Webb et al., 2023; 2025). While these models increasingly demonstrate human-like behavioral responses, alignment at the behavioral level does not logically entail alignment at the mechanistic or representational level. Indeed, task performance can decouple from neural similarity at high accuracy levels (Schrimpf et al., 2020), and models frequently achieve success via "shortcut" strategies that diverge from human internal logic (Geirhos et al., 2020). Therefore, a critical open question is whether these models also converge on human-like internal representations and computational processes. To investigate this, we extend recent work on brain–model alignment

[*]Corresponding author: c.pinier@uva.nl
[†]Shared senior authorship

in perception and language (e.g., Doerig et al., 2024; Lei et al., 2025) to abstract reasoning, comparing human behavior and neural activity to those of eight open-source LLMs on a pattern-completion task.

LLMs' ALIGNMENT WITH HUMAN BEHAVIOR AND CORTICAL PATTERNS

Deep neural networks of the 2010s achieved near-human accuracy in vision tasks (e.g., Krizhevsky et al., 2012; LeCun et al., 2015) but offered limited insight into general cognition and often failed to generalize (Bowers et al., 2022). The arrival of transformer-based LLMs marked a profound shift, both in the field of AI and in the cognitive sciences. These models have achieved unprecedented levels of generalization, enabling a broad spectrum of capabilities reminiscent of human creativity and intelligence. This rapid progress, coupled with the increasing availability of open-source LLMs, has opened up exciting research avenues enabling neuroscience-like experiments on the internal activations of these models.

Recent work demonstrates that LLM latent spaces (i.e., the internal representations emerging in hidden layers) often mirror human representational spaces. For example, GPT embeddings capture human similarity judgments of actions (Dima et al., 2024), BERT aligns with the implicit representational geometry of humans on a semantical odd-one-out task (Iaia et al., 2025), and GPT-3/4 replicate perceptual structures such as the color wheel (Marjieh et al., 2024).

Crucially, these parallels extend to cortical activity. Model activations systematically predict human brain responses in language and vision tasks (Caucheteux & King, 2022; Schrimpf et al., 2021; Goldstein et al., 2024; Mischler et al., 2024; Lei et al., 2025; Doerig et al., 2024). For instance, Schrimpf et al. (2021) found GPT-2 best predicted fMRI and ECoG signals during naturalistic language processing, while Lei et al. (2025) showed that instruction-tuned models correlate more strongly with brain data, with peak predictivity in intermediate layers (Caucheteux & King, 2022; Mischler et al., 2024).

Together, evidence across fMRI, EEG, and ECoG suggests that when LLMs achieve human-like performance, their internal activations resemble brain representational geometry. Yet most studies focus on perceptual or linguistic tasks, where low-level features may inflate scores (Feghhi et al., 2024). Abstract reasoning, which requires rule induction and relational generalization, thus offers a more stringent benchmark.

INVESTIGATING REASONING PROCESSES BETWEEN HUMANS AND AI VIA ABSTRACTIONS

As described earlier, abstract reasoning corresponds to the ability of identifying patterns, rules, and relationships from limited information and applying them to new or different contexts. It is a key element of executive functioning, allowing individuals to think about and manipulate concepts, events, and objects that are not immediately present. This capacity for higher-order thinking is tightly linked to the concept of fluid intelligence (Ferrer et al., 2009; Chuderski, 2022), which is thought to be supported by cortical networks in frontal and parietal brain regions (Caudle et al., 2023; Choi et al., 2008; Duncan, 2010; Gray et al., 2003; Perfetti et al., 2007; Santarnecchi et al., 2017; Tschentscher et al., 2017; Zurrin et al., 2024). It also appears as a crucial aspect of what distinguishes human intelligence from current artificial intelligence.

LLMs now perform strongly on many reasoning benchmarks (Wang et al., 2024), yet their abilities remain uneven: models that excel on these tests often falter on commonsense or deeper abstraction tasks (Williams & Huckle, 2024). Studies highlight both successes (Webb et al., 2023; Musker et al., 2025) and fragility, with performance dropping under small task variations (Gawin et al., 2025; Gendron et al., 2024; Hersche et al., 2024; Lee et al., 2025; Lewis & Mitchell, 2024; Li et al., 2025; Liang et al., 2025; McCoy et al., 2024; Mitchell et al., 2023; Nguyen et al., 2025; Palmarini & Mitchell, 2024; Sourati et al., 2024; Stevenson et al., 2025; Yang et al., 2025; Yax et al., 2024). Even when correct, these models often seem to rely on brittle heuristics rather than rule manipulation (Lewis & Mitchell, 2024; Lee et al., 2025; Chollet, 2019).

Although a substantial body of research has begun to examine the internal mechanisms and representational alignment of LLMs with humans, these studies often treat behavioral performance and representational similarity independently. Few studies have concurrently leveraged both dimensions to investigate whether behavioral alignment is accompanied by a corresponding representational align-

ment, particularly on reasoning tasks. Because standard EEG/fMRI methods often lock responses to experimenter-defined events, they may miss participants' strategies. We therefore leverage fixation-related potentials (FRPs), which time-lock neural signals to gaze fixations, offering a more natural window into reasoning dynamics.

USING FIXATION-RELATED POTENTIALS AS AN ECOLOGICALLY-VALID WINDOW INTO HUMAN COGNITION

Traditional cognitive neuroscience often relies on static stimuli and controlled responses (e.g., button presses), which can miss the natural dynamics of cognition during real-world tasks like reading or visual search. To improve ecological validity, we combine EEG with eye-tracking, enabling neural activity to be time-locked to self-initiated gaze fixations.

Fixation-related potentials (FRPs) are derived from this co-registration. Unlike standard stimulus-locked ERPs, FRPs reflect ongoing, self-driven cognitive processing (Degno & Liversedge, 2020). Their rapid frequency during free viewing also allows high data yield in short sessions.

Our focus on frontal EEG electrodes is motivated by extensive neuroimaging and electrophysiological literature linking the prefrontal cortex (PFC) to higher-order abstract reasoning. Evidence from multiple studies consistently identifies the frontoparietal network, and particularly the bilateral and dorsolateral PFC, as the primary neural substrate for inductive and relational reasoning (Wertheim & Ragni, 2018), with the PFC showing a distinctive capacity to represent sequential and hierarchical structures (Veselic et al., 2025). Thus, FRPs from the frontal cortex were compared to hidden-layer activations in open-source LLMs to assess representational alignment during abstract reasoning.

CURRENT STUDY

This study is structured around two objectives: i) to assess whether current open-source LLMs can perform a relatively simple reasoning task involving arbitrary symbols and abstract patterns in a manner that exhibits similar behavioral patterns to humans, ii) to explore whether the internal representations formed within these LLMs align with human cortical activity recorded via EEG.

Twenty-five participants completed an abstract-pattern-completion task while EEG and gaze data were recorded. Each trial featured a sequence of icons arranged by an implicit rule (e.g., ABCDD-CBA). Eight LLMs were given a text-based version of the same task, replacing icons with one-word labels. We compared their performance and hidden-layer activations to human behavior and EEG responses.

While previous studies have utilized representational similarity analysis (RSA) to look at alignment between LLMs and neural data in the vision and language domains (e.g., Caucheteux & King, 2022; Doerig et al., 2024), this work is, to our knowledge, the first to investigate whether such geometric convergence extends to the domain of abstract, rule-based reasoning.

2 RESULTS

COMPARISON OF LLM TO PARTICIPANT PERFORMANCE

Both human participants (N=25) and LLMs show pattern-specific difficulty, with pattern ABCDEEDC being the most difficult for both human participants and LLMs On average, humans outperform all LLMs, with an overall accuracy of 82.47% (SD = 20.38%) vs. 40.59% (SD = 33.08%). However, the $\sim$ 70 billion parameter models, namely, Qwen2.5-72B, Deepseek-R1-Distill-Llama-70B, and Llama-3.3-70B, differentiate themselves from the rest with accuracy scores between 75.00% and 81.75% (compared to less than 40% for all the others). Ideal LLM candidates for modelling human cognition on this task should display both a high score on the task and a high correlation with human accuracy profile (top right quadrant of Figure 1). That is, their performance by pattern type should closely follow that of humans. Two of them, Qwen2.5-72B and Deepseek-R1-Distill-Llama-70B, fit this description with an overall accuracy of 80.50% and 75.00%, and a Pearson correlation with human performance of .71 and .70, respectively. Surprisingly, Llama-3.3-70B, despite being the most accurate LLM (M = 81.75%, SD = 38.67%), aligns poorly with the human accuracy profile (r = .27). On the other hand, Phi-4 shows the opposite trend, with an overall

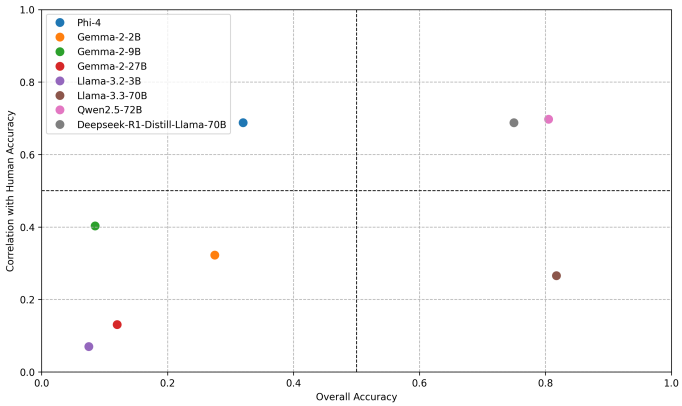
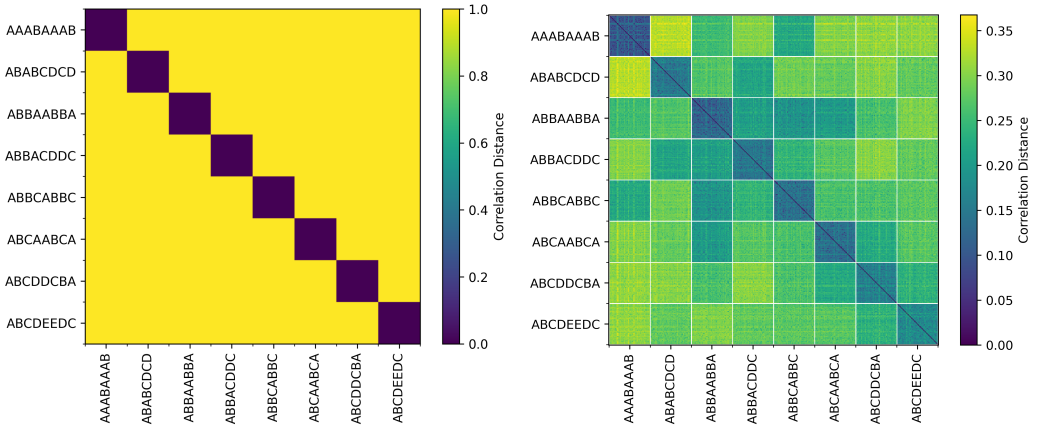


Figure 1: LLMs' overall task accuracy vs. correlation with human accuracy profile. Each dot represents one LLM and is positioned by its overall accuracy on the 400 abstract-sequence trials (x-axis) and the Pearson correlation (r) between its 8-pattern accuracy vector and that of human participants (y-axis).

low accuracy on the task but a better fit to the human response structure, with a correlation value (r = .67) on par with that of Qwen2.5-72B and Deepseek-R1-Distill-Llama-70B.



(a) (b)

Figure 2: Trial-level representational dissimilarity matrices (RDMs) generated from 400 unique trials (50 trials per pattern type). **Left:** Task-optimal Reference RDM, encoding perfect within-pattern similarity and maximal between-pattern dissimilarity; **Right**: grand-average LLM RDM, computed from the activations of each LLM's task-optimal layer; brighter colors denote greater dissimilarity.

IDENTIFICATION OF TASK-OPTIMAL LLM LAYER

Having established which models not only score on par with humans, but also mirror their accuracy profile, we next asked where in each LLM's architecture do the eight abstract pattern classes become most cleanly separated. To find that locus, we used representational similarity analysis (RSA; Kriegeskorte et al., 2008). We built a trial-level (400 × 400) representational dissimilarity matrix (RDM) that encodes perfect within-pattern similarity (0 on the eight 50 × 50 diagonal blocks) and maximal between-pattern dissimilarity (1 elsewhere; Figure 2(a)). This design RDM encodes the same pattern type no matter the individual icon/words in the pattern, which varied per experimental trial in both humans (with icons) and LLMs (with words). We then correlated this matrix with the RDM produced by each layer of each LLM.

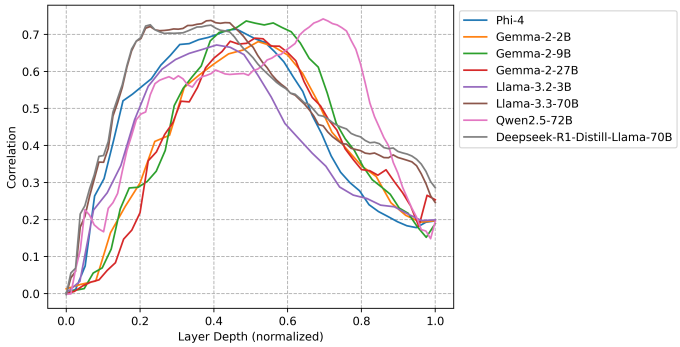


Figure 3: Layer-wise Pearson correlation between LLM layer RDMs and reference RDM.

We found that layer-wise correlations to this reference RDM form a pronounced inverted-U profile for each LLM (Figure 3): similarity is negligible in the early layers, but then rises steadily and reaches its apex in the intermediate blocks — on average at 47.52% (SD = 13.17%) of the layer depth — before tapering off toward the output layer. These "sweet-spot" layers reveal a markedly block-structured RDM (Figure 2(b)), indicating that internal representations, in this zone, cluster trials almost perfectly by abstract pattern class, making these patterns an explicit organizing axis of the latent space. Notably, the strength of this task-specific geometry is related to LLM behavioral competence: the Pearson correlation between a model's overall accuracy and its maximum similarity to the reference RDM is r = .71 (SD = .03). The layer that maximized this correlation was designated as the model's "task-optimal" layer, and its corresponding RDM was further used for direct comparisons with human neural data. These results confirm that LLMs which perform better also carve the most distinct pattern clusters in their internal representations. After isolating the task-optimal layer in each LLM, we averaged its activations over the trials belonging to each abstract pattern type, producing eight vectors that were then used to build pattern-level RDMs.

LLM LAYER COMPARISON TO HUMAN EEG DATA

To test whether the representations of the previously identified task-optimal layers resonate with human neural signals, we ran a second RSA in which each LLM's RDM was compared with three EEG RDMs constructed from frontal-electrode data and averaged across participants:

1. FRPs were obtained by first averaging the EEG signal across all icon fixations within a trial. These measures were thought to capture the aggregate, self-paced processing that unfolds as people inspect the abstract sequence of icons.
2. Response-locked ERPs, corresponding to a more "classical" approach — for example, response-locked ERPs in central electrodes are thought to reflect evidence accumulation during decision-making (Gluth et al., 2013; Lui et al., 2021) — were time-locked to button presses (or to the end of the response window).
3. Resting EEG activity was extracted from inter-trial intervals to provide a cognitive "null" baseline.

For each of these three group-level datasets, we once again pooled the trials belonging to the same abstract pattern type, averaged them into eight pattern-wise vectors, and used these to construct the corresponding pattern-level RDMs.

Group 1	**Group 2**	**Mean diff**	**p-adj**	**lower**	**upper**	**reject**
Response ERP	FRP	0.26	< 0.01	0.22	0.30	True
Resting EEG	FRP	0.30	< 0.01	0.26	0.34	True
Resting EEG	Response ERP	0.04	0.07	-0.0025	0.08	False

Table 2: Multiple Comparison of Means - Tukey HSD, family-wise error rate (FWER)=0.05

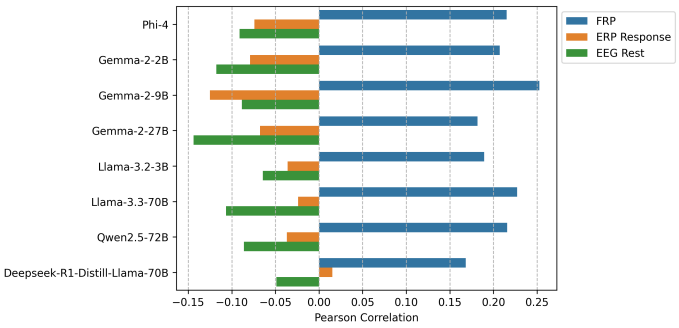


Figure 4: Pattern-level RSA between Human group-averaged RDMs and LLMs' task-optimal layer RDMs

	Correlation (r)		**p-value**	
Type	mean ± SD	range	mean ± SD	range
FRP	0.21 ± 0.03	0.17 – 0.25	0.16 ± 0.03	0.12 – 0.20
ERP Response	-0.05 ± 0.04	-0.13 – 0.02	0.56 ± 0.08	0.45 – 0.69
EEG Rest	-0.09 ± 0.03	-0.14 – -0.05	0.67 ± 0.05	0.59 – 0.75

r = Pearson correlation; p values from 10 000-iteration permutation test.

Table 1: Pattern-level RSA correlations and permutation-test p-values

Although none of the LLM-EEG correlations reached permutation significance (all p $>$ .05; see Table 1 and *Limitations* section), the magnitudes and p-values of the FRP data diverged systematically from the response-ERP and resting EEG data (see Figure 4 and Figure 8). FRP similarities were uniformly positive, spanning r $\approx$ .17–.25 (M = .21, SD = .03, M p-val = .16, SD p-val = .03), whereas response-locked ERPs were negative or hovered near zero, r $\approx$ –.125 – .01 (M = -.05, SD = .04, M p-val = .56, SD p-val =.08). Resting ERPs were consistently negative, r $\approx$ –.15–.05 (M = -.09, SD = .03, M p-val = .67, SD p-val = .05). A post-hoc Tukey HSD test on the correlation coefficients confirmed that the FRP correlations were significantly larger than those obtained from either of the other two EEG measures (Table 2). Specifically, FRPs exceeded resting-EEG correlations by an average of .30 (95 % CI [.26, .34], p-adj $<$ .01) and response-locked ERP correlations by .26 (95 % CI [.22, .30], p-adj $<$ .01).

Thus, only gaze-linked EEG data (FRPs) — constructed by averaging all fixations within a trial — seem to potentially carry any detectable trace of the same abstract-pattern geometry encoded in the LLMs' mid-layers, while response-locked or resting EEG data do not. These modest yet systematic FRP correlations (r $\approx$ .17–.25) complement an earlier result from the LLMs: across the eight models, the more distinctly a mid-layer encoded the eight pattern categories, the higher the model's overall accuracy (r = 0.71). In other words, the representations that make a model succeed on the task are the same representations that seem to faintly reappear in human frontal FRPs, hinting at a shared coding of abstract patterns.

3 Discussion

In this work, we compared the behavior and neural representations of human participants to those of eight open-source LLMs on an abstract-pattern-completion task. We show that only the largest models ($\sim$ 70 billion parameters each) approach humans' overall performance with a mean accuracy of 79.08% vs. 82.47% for human participants. Looking at accuracy profiles, Qwen2.5-72B, Deepseek-R1-Distill-Llama-70B, and Phi-4 show the best alignment with humans. A particularly interesting comparison comes from Llama-3.3-70B and its derivative DeepSeek-R1-Distill-Llama-70B. Both share the same 70-billion-parameter transformer backbone, yet differ in their second-stage training. The base model, Llama-3.3-70B, relies solely on large-scale next-token prediction, whereas DeepSeek-R1 is distilled (i.e., trained to imitate a teacher model's chain-of-thought outputs on a

curated dataset) and then fine-tuned with reinforcement learning so that it is encouraged to consistently produce those explicit reasoning steps. This procedural change produces a clear trade-off: in comparison to Llama-3.3-70B, the reasoning-optimised variant trades $\sim$7 percentage-points of accuracy (75.00% vs 81.75%) for a 2.6-fold increase in human-likeness, as measured by Pearson's r on accuracy by pattern type (.27 vs. .70). Encouraging step-by-step reasoning might therefore bring about more human-like error-patterns, albeit at the cost of a modest reduction in overall capabilities.

Second, across architectures, layer-wise correlations to a design RDM that encodes the abstract pattern categories (maximal within-pattern similarity and maximal between-pattern dissimilarity) trace a similar inverted-U alignment profile, reaching a peak in intermediate layers (Figure 3). These correspondences were positively associated with overall task accuracy, implying that better performers encode the task's structure more strongly. Taken together, these results suggest that mid-level representations most faithfully capture the structure of the task, which echoes previous findings highlighting the importance of intermediate layers in various deep learning architectures (Ju et al., 2024; Lei & Cooper, 2025; Meng et al., 2023; Park & Kim, 2025; Skean et al., 2024; 2025; Sun et al., 2025; Zhang et al., 2024). Zhang et al. (2025), for example, demonstrated the existence of a handful of "cornerstone" layers, typically in the early-to-mid section, carrying a dominant share of task-relevant information, while results from Meng et al. (2023) showed that mid-layers in feed-forward modules are crucial for storing factual knowledge. RDMs derived from the mid-level layers of every model converged on a remarkably similar geometry that clusters the different trials into their abstract pattern classes (see Figure 2(b) and Figure 7), which is also corroborated by earlier findings (Lan et al., 2025; Wolfram & Schein, 2025; Zhang et al., 2024) and the Platonic Representation Hypothesis. The latter states that "neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces." (Huh et al., 2024).

Third, we show moderate correlations (r $\approx$.17–.25) in representational geometry between task-optimal LLM layers and human cortical data extracted from participants' frontal FRPs. Although the permutation test fell short of statistical significance, there was a consistent trend showing an increase in correlation scores and decrease in their associated p-values for the FRP data in comparison with the other two EEG datasets, built from response-locked ERPs and resting EEG activity. EEG inevitably imposes a modest signal-to-noise ceiling, and our current FRP analysis might not be able to capture all or most of the reasoning-related cognitive activity present in the cortical data. However, if we consider that the LLMs capture the inherent logic of the task in their internal representations, and that FRP RDMs correlate consistently more strongly with the "task-optimal" layers' RDMs than the response-lock and resting activity RDMs, then we might conclude that abstract reasoning activity in the human brain can be at least partially mimicked by LLMs' representations. Crucially, the cortical alignment we observe at intermediate LLM layers mirrors a converging pattern in the literature: mid-layer activations were already shown to best predict brain responses during sentence processing in various brain recording modalities (Caucheteux & King, 2022; Lei et al., 2025; Mischler et al., 2024).

Limitations and future directions

This study offers initial evidence for LLM–brain alignment during abstract reasoning, but several limitations remain. First, the small EEG sample (N = 25) limits statistical power; larger cohorts would improve robustness and allow individual-difference analyses. Second, the task modality differed across agents: humans solved a visuospatial puzzle, while LLMs processed text, potentially weakening alignment. Third, our use of representational similarity analysis reveals *where* abstract-rule information emerges but not *how* it arises. Combining RSA with causal interpretability tools (e.g., activation patching, attention ablation) may uncover underlying mechanisms and generalization. Fourth, while FRPs showed the strongest correspondence with LLM activations, results did not reach significance—possibly due to EEG noise and the conservativeness of permutation testing on low-dimensional RDMs (Nili et al., 2014). Fifth, future work could use fMRI to provide spatially detailed comparisons. Lastly, further analysis of eye-tracking data—such as comparing fixation heatmaps with model attention—could reveal shared attentional dynamics.

4 METHODS

TASK DESIGN

In each trial, participants viewed a sequence of icons arranged according to a specific underlying pattern (e.g., "ABBACDDC"). Eight unique patterns were used in the experiment:

AAABAAAB	ABBACDDC	ABCDDCBA	ABABCDCD
ABBCABBC	ABCDEEDC	ABBAABBA	ABCAABCA

For each trial, the final icon in the sequence was replaced by a question mark. Participants were then required to select the icon from a set of four multiple-choice options that correctly continued the sequence. Participants responded by pressing one of four computer keyboard buttons corresponding to each response. This design was adapted to a text-based version for LLMs, which were presented with sequences of one-word labels describing the icons.

In the lab version of the experiment, participants were seated at about 60 cm from a computer monitor with their head stabilized on an adjustable chin rest to minimize movement, while EEG signals were recorded via a 64-electrode cap and gaze data were captured simultaneously with an eye tracker. The icons of a sequence were first individually presented for 600 ms in a randomized order (but at their respective location in the sequence) on the top part of the screen. This "encoding phase" ensured that participants could register the visual features of every icon without yet attempting to solve the pattern. Subsequently, four choice icons were displayed in a similar manner, but on the lower part of the screen. After these separate presentations, the entire sequence and the four options were presented simultaneously on the screen and remained until the participant responded by pressing a key or until a maximum duration of 12 seconds was reached. We refer to this second phase as the "decision phase". By temporally separating these two phases, we aimed to isolate the neural signals related to reasoning processes from those associated with more basic perceptual processing.

The full experiment consisted of 400 unique trials divided into 5 sessions, with 50 trials per pattern type. Each session was divided into 4 blocks of 20 trials, for a total of 80 trials with 10 trials per pattern type, presented in random order.

PARTICIPANTS

Twenty-five healthy adults were recruited from the University's participant pool through an online advertisement. The advertisement specified the eligibility criteria ($\geq$ 18 years; normal or corrected-to-normal vision; no personal or familial history of epilepsy) and the study logistics: five sessions of about one to two hours each. Volunteers were offered either €100 or six course-credit units as compensation for the completion of the whole study. Five participants withdrew before completing the whole experiment. Partial datasets from these were retained and included in all analyses. All participants gave written informed consent, and the protocol was approved by the Faculty Ethics Review Board.

LLMS

Eight open-source LLMs were downloaded and run locally using the Hugging Face library: Phi-4, Gemma-2-2B, Gemma-2-9B, Gemma-2-27B, Llama-3.2-3B, Llama-3.3-70B, Qwen2.5-72B, Deepseek-R1-Distill-Llama-70B The models were queried 400 different times, that is, one trial at a time with no context carried over between prompts (i.e., no memory of previous trials), and each of their layers' outputs were extracted.

EEG & EYE TRACKING APPARATUS

EEG data were collected using a 64-electrode headcap from BioSemi, arranged according to the international 10-20 system and connected to an EEG amplifier system with a sampling rate of 2048 Hz. Eye movements were recorded from the participants' dominant eye using EyeLink 1000 Plus system with a sampling rate of 2000 Hz.

DATA ANALYSIS

BEHAVIORAL ANALYSIS

For the behavioral analysis, we first computed each participant's mean accuracy across all 400 trials and their mean accuracy within each of the eight abstract-pattern categories. The same two metrics were calculated for every LLM. To assess how closely a model's accuracy profile matched human behavior, we then correlated its pattern-accuracy vector with the group-average human vector using Pearson correlation.

LLMS' LAYER ACTIVATIONS ON SEQUENCE TOKENS

For each trial, we extracted hidden-state activations from every layer of each LLM, but retained only those corresponding to the tokens in the abstract sequence (e.g., "star star star guitar ... ?"). Activations for tokens corresponding to the rest of the prompt were discarded. The resulting trial-level vectors thus capture each model's internal representations of the abstract sequence itself.

REPRESENTATIONAL SIMILARITY ANALYSIS (RSA)

FRP Dataset. To construct representational dissimilarity matrices (RDMs) from the human EEG data, FRPs were extracted from 17 frontal electrodes (see Figure 6):

1. Gaze fixations were automatically identified and labeled by the Eyelink 1000 Plus software. We extracted fixations from the decision phase (see Task design section in the Methods) and only kept those that landed on either of the eight icons forming the abstract sequence (top row of Figure 5), disregarding those that landed on the choice options (bottom row of Figure 5).
2. For every retained fixation, we epoched the concurrently recorded EEG from 0 – 600 ms relative to fixation onset, as free-viewing fixations usually last between 200 and 500 ms (Engbert, 2006).
3. Within a trial, the fixation-locked waveforms were averaged together, yielding one 600-ms FRP per trial that we interpret as a composite neural trace of the abstract pattern being inspected.

Additional EEG measures. Two complementary EEG measures were derived from the same 17 frontal electrodes:

i) Response-locked ERP: Epochs spanned -1000 to 0 ms around the button press that completed a trial. No baseline or further processing were applied.
ii) Resting EEG (control): To obtain a cognitive "null" baseline, we extracted epochs from the inter-trial period, during which a fixation cross was displayed at the center of the screen. More specifically, we selected data from -1000 to 0 ms around the start of a trial, which was defined as the 1-second mark before the appearance of the first stimulus on screen, after a variable inter-trial period of 1, 2, or 3 seconds. At that instant the fixation cross was still onscreen and no task-relevant stimuli had yet been presented. Again, no baseline or further processing were applied.

Human RDMs. For each EEG measure, the 400 trial-wise waveforms of a participant were first vectorised along the time and electrodes dimensions, resulting in a matrix of 400 (trials) $\times$ 20,893 (17 electrodes $\times$ 1,229 timepoints). This data was averaged across participants to form group-level datasets. Pattern-level averages were then obtained from the group-level datasets by collapsing over the 50 trials belonging to each pattern type, resulting in 8 $\times$ 20,893 matrices. Finally, we constructed one pattern-level RDM (8 × 8 matrix) per EEG measure, capturing the pair-wise dissimilarities among the eight pattern types from the group-level data. Correlation distance was used as the dissimilarity metric, which quantifies the dissimilarity between two patterns as 1 - r based on the Pearson correlation between their vectors.

LLM RDMs. A 400 $\times$ 400 trial-level RDM was built from the hidden layers activations of each LLM (see *LLMs' layer activations on sequence tokens* section) by computing correlation distance between all pairs of activation vectors. To locate the layer that best captured the task's abstract structure, we correlated each layer's RDM (Pearson's r) with the reference RDM (see Figure 2(a)), which assigns 0 to within-pattern pairs and 1 to between-pattern pairs. For each LLM, the layer with

the highest correlation was designated as that model's task-optimal layer. Finally, from every task-optimal layer, we averaged the activation vectors across pattern type and computed their pair-wise correlation distances to yield an 8×8 pattern-level RDM. These task-optimal, pattern-level RDMs served as the basis for all subsequent comparisons with the human EEG RDMs.

RDM Comparison. To assess the similarity between the representational geometries of human EEG responses and LLM activations, we compared the respective pattern-level RDMs using RSA. The similarity between two RDMs was quantified using Pearson correlation. A high correlation indicates that the two RDMs share a similar representational geometry, suggesting that the underlying patterns of activity are organized in a comparable fashion.

Permutation testing. To evaluate the statistical significance of the observed correlation, we employed a non-parametric permutation test with 10,000 iterations. For each iteration, the human FRP dataset was disrupted by randomly permuting the condition indices. The FRP RDM was recomputed from the permuted data and then compared to the fixed LLM RDMs using the same similarity metric (Pearson correlation). A null distribution of similarity scores was generated from these permutations, and the p-value was calculated as the fraction of permuted correlations whose magnitude exceeded that of the observed similarity.

5 Open-access Statement

All data and code required to reproduce the study's findings are openly available. The anonymized behavioral, EEG, and eye-tracking datasets, as well as the LLM responses and activations, are available on Figshare (`https://doi.org/10.21942/uva.29573534`). The analysis pipeline, figure generation scripts, and stimuli are available on GitHub (`https://github.com/chris-pinier/abstract_reasoning`).

6 Acknowledgments

We are appreciative of the Dutch government and the University of Amsterdam for providing a starting grant that funds joint work by Christopher Pinier, Claire E. Stevenson, and Michael D. Nunez. This research was also funded in part by 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 E. Stevenson.

References

Jeffrey S. Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E. Hummel, Rachel F. Heaton, Benjamin D. Evans, Jeffrey Mitchell, and Ryan Blything. Deep Problems with Neural Network Models of Human Vision. *The Behavioral and Brain Sciences*, pp. 1–74, December 2022. ISSN 1469-1825. doi: 10.1017/S0140525X22002813.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: early experiments with GPT-4, April 2023. URL `http://arxiv.org/abs/2303.12712`. arXiv:2303.12712 [cs].

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):1–10, February 2022. ISSN 2399-3642. doi: 10.1038/s42003-022-03036-1. URL `https://www.nature.com/articles/s42003-022-03036-1`. Publisher: Nature Publishing Group.

M. M. Caudle, A. D. Spadoni, D. M. Schiehser, A. N. Simmons, and J. Bomyea. Neural activity and network analysis for understanding reasoning using the matrix reasoning task. *Cognitive Processing*, 24(4):585–594, 2023. ISSN 1612-4782. doi: 10.1007/s10339-023-01152-2. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10533635/`.

Yu Yong Choi, Noah A. Shamosh, Sun Hee Cho, Colin G. DeYoung, Min Joo Lee, Jong-Min Lee, Sun I. Kim, Zang-Hee Cho, Kyungjin Kim, Jeremy R. Gray, and Kun Ho Lee. Multiple bases of human intelligence revealed by cortical thickness and neural activation. *Journal of Neuroscience*, 28(41):10323–10329, October 2008. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.3259-08.2008. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6671030/.

François Chollet. On the measure of intelligence, November 2019. URL http://arxiv.org/abs/1911.01547. arXiv:1911.01547 [cs].

Adam Chuderski. Fluid intelligence emerges from representing relations. *Journal of Intelligence*, 10(3):51, August 2022. ISSN 2079-3200. doi: 10.3390/jintelligence10030051. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9396997/.

Federica Degno and Simon P. Liversedge. Eye movements and fixation-related potentials in reading: a review. *Vision*, 4(1):11, February 2020. ISSN 2411-5150. doi: 10.3390/vision4010011. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7157570/.

Diana C. Dima, Sugitha Janarthanan, Jody C. Culham, and Yalda Mohsenzadeh. Shared representations of human actions across vision and language. *Neuropsychologia*, 202:108962, September 2024. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2024.108962. URL https://www.sciencedirect.com/science/article/pii/S0028393224001775.

Adrien Doerig, Tim C. Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. Visual representations in the human brain are aligned with large language models, July 2024. URL http://arxiv.org/abs/2209.11737. arXiv:2209.11737 version: 2.

John Duncan. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, 14(4):172–179, April 2010. ISSN 1879-307X. doi: 10.1016/j.tics.2010.01.004.

Ralf Engbert. Microsaccades: a microcosm for research on oculomotor control, attention, and visual perception. In S. Martinez-Conde, S. L. Macknik, L. M. Martinez, J. M. Alonso, and P. U. Tse (eds.), *Progress in Brain Research*, volume 154 of *Visual Perception*, pp. 177–192. Elsevier, January 2006. doi: 10.1016/S0079-6123(06)54009-9. URL https://www.sciencedirect.com/science/article/pii/S0079612306540099.

Ebrahim Feghhi, Nima Hadidi, Bryan Song, Idan A. Blank, and Jonathan C. Kao. What Are Large Language Models Mapping to in the Brain? A Case Against Over-Reliance on Brain Scores, June 2024. URL http://arxiv.org/abs/2406.01538. arXiv:2406.01538 [cs] version: 1.

Emilio Ferrer, Elizabeth D. O'Hare, and Silvia A. Bunge. Fluid reasoning and the developing brain. *Frontiers in Neuroscience*, 3(1):46–51, May 2009. ISSN 1662-453X. doi: 10.3389/neuro.01.003.2009.

Cole Gawin, Yidan Sun, and Mayank Kejriwal. Navigating semantic relations: challenges for language models in abstract common-sense reasoning, February 2025. URL http://arxiv.org/abs/2502.14086. arXiv:2502.14086 [cs].

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL http://arxiv.org/abs/2004.07780. arXiv:2004.07780 [cs].

Gaël Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. Large language models are not strong abstract reasoners, January 2024. URL http://arxiv.org/abs/2305.19555. arXiv:2305.19555 [cs].

Sebastian Gluth, Jörg Rieskamp, and Christian Büchel. Classic EEG motor potentials track the emergence of value-based decisions. *Neuroimage*, 79:394–403, October 2013. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2013.05.005.

Ariel Goldstein, Avigail Grinstein-Dabush, Mariano Schain, Haocheng Wang, Zhuoqiao Hong, Bobbi Aubrey, Mariano Schain, Samuel A. Nastase, Zaid Zada, Eric Ham, Amir Feder, Harshvardhan Gazula, Eliav Buchnik, Werner Doyle, Sasha Devore, Patricia Dugan, Roi Reichart, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Orrin Devinsky, Adeen Flinker, and Uri Hasson. Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications*, 15(1):2768, March 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-46631-y. URL https://www.nature.com/articles/s41467-024-46631-y. Publisher: Nature Publishing Group.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. MEG and EEG data analysis with MNE-python. *Frontiers in Neuroscience*, 7:267, December 2013. ISSN 1662-4548. doi: 10.3389/fnins.2013.00267.

Jeremy R. Gray, Christopher F. Chabris, and Todd S. Braver. Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, 6(3):316–322, March 2003. ISSN 1546-1726. doi: 10.1038/nn1014. URL https://www.nature.com/articles/nn1014. Publisher: Nature Publishing Group.

Michael Hersche, Giacomo Camposampiero, Roger Wattenhofer, Abu Sebastian, and Abbas Rahimi. Towards learning to reason: comparing LLMs with neuro-symbolic on arithmetic relations in abstract reasoning, December 2024. URL http://arxiv.org/abs/2412.05586. arXiv:2412.05586 [cs] version: 1.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, July 2024. URL http://arxiv.org/abs/2405.07987. arXiv:2405.07987 [cs].

Cosimo Iaia, Bhavin Choksi, Emily Wiebers, Gemma Roig, and Christian J. Fiebach. The representational alignment between humans and language models is implicitly driven by a concreteness effect, May 2025. URL http://arxiv.org/abs/2505.15682. arXiv:2505.15682 [cs] version: 1.

Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. How large language models encode context knowledge? A layer-wise probing study, March 2024. URL http://arxiv.org/abs/2402.16061. arXiv:2402.16061 [cs].

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, November 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008/full. Publisher: Frontiers.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Quantifying feature space universality across large language models via sparse autoencoders, May 2025. URL http://arxiv.org/abs/2410.06981. arXiv:2410.06981 [cs] version: 4.

Dominic Langlois, Sylvain Chartier, and Dominique Gosselin. An introduction to independent component analysis: InfoMax and FastICA algorithms. *Tutorials in Quantitative Methods for Psychology*, 6(1):31–38, March 2010. ISSN 1913-4126. doi: 10.20982/tqmp.06.1.p031. URL http://www.tqmp.org/RegularArticles/vol06-1/p031.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL https://www.nature.com/articles/nature14539. Publisher: Nature Publishing Group.

Seungpil Lee, Woochang Sim, Donghyeon Shin, Wongyu Seo, Jiwon Park, Seokki Lee, Sanha Hwang, Sejin Kim, and Sundong Kim. Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. *ACM Trans. Intell. Syst. Technol.*, pp. 3712701, January 2025. ISSN 2157-6904. doi: 10.1145/3712701. URL https://dl.acm.org/doi/10.1145/3712701. Just Accepted.

Ge Lei and Samuel J. Cooper. The representation and recall of interwoven structured knowledge in LLMs: a geometric and layered analysis, February 2025. URL http://arxiv.org/abs/2502.10871. arXiv:2502.10871 [cs] version: 1.

Yu Lei, Xingyang Ge, Yi Zhang, Yiming Yang, and Bolei Ma. Do large language models think like the brain? Sentence-level evidence from fMRI and hierarchical embeddings, May 2025. URL http://arxiv.org/abs/2505.22563. arXiv:2505.22563 [cs].

Martha Lewis and Melanie Mitchell. Evaluating the robustness of analogical reasoning in large language models, November 2024. URL http://arxiv.org/abs/2411.14215. arXiv:2411.14215 [cs].

Adam Li, Jacob Feitelberg, Anand Prakash Saini, Richard Höchenberger, and Mathieu Scheltienne. MNE-ICALabel: automatically annotating ICA components with ICLabel in python. *Journal of Open Source Software*, 7(76):4484, August 2022. ISSN 2475-9066. doi: 10.21105/joss.04484. URL https://joss.theoj.org/papers/10.21105/joss.04484.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D. Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, and Hokin Deng. Core knowledge deficits in multi-modal language models, June 2025. URL http://arxiv.org/abs/2410.10855. arXiv:2410.10855 [cs].

Shanchao Liang, Spandan Garg, and Roshanak Zilouchian Moghaddam. The SWE-bench illusion: when state-of-the-art LLMs remember instead of reason, June 2025. URL http://arxiv.org/abs/2506.12286. arXiv:2506.12286 [cs].

Kitty K. Lui, Michael D. Nunez, Jessica M. Cassidy, Joachim Vandekerckhove, Steven C. Cramer, and Ramesh Srinivasan. Timing of readiness potentials reflect a decision-making process in the human brain. *Computational Brain & Behavior*, 4(3):264–283, September 2021. ISSN 2522-087X. doi: 10.1007/s42113-020-00097-5.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445, September 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-72071-1. URL https://www.nature.com/articles/s41598-024-72071-1. Publisher: Nature Publishing Group.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, October 2024. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2322420121. URL https://www.pnas.org/doi/10.1073/pnas.2322420121. Publisher: Proceedings of the National Academy of Sciences.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT, January 2023. URL http://arxiv.org/abs/2202.05262. arXiv:2202.05262 [cs].

Gavin Mischler, Yinghao Aaron Li, Stephan Bickel, Ashesh D. Mehta, and Nima Mesgarani. Contextual Feature Extraction Hierarchies Converge in Large Language Models and the Brain, January 2024. URL https://arxiv.org/abs/2401.17671v1.

Melanie Mitchell, Alessandro B. Palmarini, and Arseny Moskvichev. Comparing humans, GPT-4, and GPT-4V on abstraction and reasoning tasks, December 2023. URL http://arxiv.org/abs/2311.09247. arXiv:2311.09247 [cs].

Sam Musker, Alex Duchnowski, Raphaël Millière, and Ellie Pavlick. LLMs as models for analogical reasoning, March 2025. URL http://arxiv.org/abs/2406.13803. arXiv:2406.13803 [cs] version: 2.

A. Newell and H. Simon. The logic theory machine–a complex information processing system. *IRE Transactions on Information Theory*, 2(3):61–79, September 1956. ISSN 2168-2712. doi: 10.1109/TIT.1956.1056797. URL https://ieeexplore.ieee.org/document/1056797.

Allen Newell. The chess machine: an example of dealing with a complex task by adaptation. In *Proceedings of the March 1-3, 1955, Western Joint Computer Conference*, AFIPS '55 (Western), pp. 101–108, New York, NY, USA, March 1955. Association for Computing Machinery. ISBN 978-1-4503-7856-7. doi: 10.1145/1455292.1455312. URL https://dl.acm.org/doi/10.1145/1455292.1455312.

Tuan Dung Nguyen, Duncan J. Watts, and Mark E. Whiting. Empirically evaluating commonsense intelligence in large language models with large-scale human judgments, May 2025. URL http://arxiv.org/abs/2505.10309. arXiv:2505.10309 [cs].

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014.

Alessandro B Palmarini and Melanie Mitchell. Abstract understanding of core-knowledge concepts: Humans vs. llms. In *ICML 2024 Workshop on LLMs and Cognition*, 2024. URL https://openreview.net/pdf?id=bFWBD4UvUk.

Hancheol Park and Geonmin Kim. Where do LLMs encode the knowledge to assess the ambiguity? In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Kareem Darwish, and Apoorv Agarwal (eds.), *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 445–452, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-industry.38/.

Bernardo Perfetti, Aristide Saggino, Antonio Ferretti, Massimo Caulo, Gian Luca Romani, and Marco Onofrj. Differential patterns of cortical activation as a function of fluid reasoning complexity. *Human Brain Mapping*, 30(2):497–510, December 2007. ISSN 1065-9471. doi: 10.1002/hbm.20519. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6871137/.

F. Perrin, J. Pernier, O. Bertrand, and J. F. Echallier. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, 72(2):184–187, February 1989. ISSN 0013-4694. doi: 10.1016/0013-4694(89)90180-6. URL https://www.sciencedirect.com/science/article/pii/0013469489901806.

Emiliano Santarnecchi, Alexandra Emmendorfer, and Alvaro Pascual-Leone. Dissecting the parieto-frontal correlates of fluid intelligence: a comprehensive ALE meta-analysis study. *Intelligence*, 63:9–28, July 2017. ISSN 0160-2896. doi: 10.1016/j.intell.2017.04.008. URL https://www.sciencedirect.com/science/article/pii/S0160289617300090.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?, January 2020. URL https://www.biorxiv.org/content/10.1101/407007v2. Pages: 407007 Section: New Results.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 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, November 2021. ISSN 1091-6490. doi: 10.1073/pnas.2105646118.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go

with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016. ISSN 1476-4687. doi: 10.1038/nature16961. URL https://www.nature.com/articles/nature16961. Publisher: Nature Publishing Group.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, October 2017. ISSN 1476-4687. doi: 10.1038/nature24270. URL https://www.nature.com/articles/nature24270. Publisher: Nature Publishing Group.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, December 2018. doi: 10.1126/science.aar6404. URL https://www.science.org/doi/10.1126/science.aar6404. Publisher: American Association for the Advancement of Science.

Oscar Skean, Md Rifat Arefin, Yann LeCun, and Ravid Shwartz-Ziv. Does representation matter? Exploring intermediate layers in large language models, December 2024. URL http://arxiv.org/abs/2412.09563. arXiv:2412.09563 [cs].

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: uncovering hidden representations in language models, February 2025. URL http://arxiv.org/abs/2502.02013. arXiv:2502.02013 [cs].

Zhivar Sourati, Filip Ilievski, Pia Sommerauer, and Yifan Jiang. ARN: analogical reasoning on narratives, September 2024. URL http://arxiv.org/abs/2310.00996. arXiv:2310.00996 [cs].

Claire E. Stevenson, Alexandra Pafford, Han L. J. van der Maas, and Melanie Mitchell. Can large language models generalize analogy solving like people can?, March 2025. URL http://arxiv.org/abs/2411.02348. arXiv:2411.02348 [cs].

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models, February 2025. URL http://arxiv.org/abs/2502.05795. arXiv:2502.05795 [cs].

Nadja Tschentscher, Daniel Mitchell, and John Duncan. Fluid intelligence predicts novel rule implementation in a distributed frontoparietal control network. *Journal of Neuroscience*, 37(18):4841–4847, May 2017. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2478-16.2017. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5426573/.

Sebastijan Veselic, Elena Gutierrez, Mohamady El-Gaby, Sandra Reinert, and Mathias Sablé-Meyer. Cognitive maps in the prefrontal cortex. *Journal of Neuroscience*, 45(46), November 2025. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1358-25.2025. URL https://www.jneurosci.org/content/45/46/e1358252025.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (MLLMs): a comprehensive survey on emerging trends in multimodal reasoning, January 2024. URL http://arxiv.org/abs/2401.06805. arXiv:2401.06805 [cs].

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent Analogical Reasoning in Large Language Models. August 2023. doi: 10.48550/arXiv.2212.09196. URL http://arxiv.org/abs/2212.09196.

Taylor W Webb, Keith J Holyoak, and Hongjing Lu. Evidence from counterfactual tasks supports emergent analogical reasoning in large language models. *PNAS Nexus*, 4(5):pgaf135, May 2025. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgaf135. URL https://doi.org/10.1093/pnasnexus/pgaf135.

Julia Wertheim and Marco Ragni. The neural correlates of relational reasoning: a meta-analysis of 47 functional magnetic resonance studies. *Journal of Cognitive Neuroscience*, 30(11):1734–1748, November 2018. ISSN 0898-929X. doi: 10.1162/jocn_a_01311. URL https://doi.org/10.1162/jocn_a_01311.

Sean Williams and James Huckle. Easy problems that LLMs get wrong, June 2024. URL http://arxiv.org/abs/2405.19616. arXiv:2405.19616 [cs].

Christopher Wolfram and Aaron Schein. Layers at similar depths generate similar activations across LLM architectures, May 2025. URL http://arxiv.org/abs/2504.08775. arXiv:2504.08775 [cs].

Yue Yang, MingKang Chen, Qihua Liu, Mengkang Hu, Qiguang Chen, Gengrui Zhang, Shuyue Hu, Guangtao Zhai, Yu Qiao, Yu Wang, Wenqi Shao, and Ping Luo. Truly assessing fluid intelligence of large language models through dynamic reasoning evaluation, June 2025. URL http://arxiv.org/abs/2506.02648. arXiv:2506.02648 [cs] version: 1.

Nicolas Yax, Hernán Anlló, and Stefano Palminteri. Studying and improving reasoning in humans and machines. *Communications Psychology*, 2(1):51, June 2024. ISSN 2731-9121. doi: 10.1038/s44271-024-00091-8. URL https://www.nature.com/articles/s44271-024-00091-8. Publisher: Nature Publishing Group.

Yang Zhang, Yanfei Dong, and Kenji Kawaguchi. Investigating layer importance in large language models. In *Proceedings of the 7th Blackboxnlp Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 469–479, Miami, Florida, US, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.29. URL https://aclanthology.org/2024.blackboxnlp-1.29.

Zhejun Zhang, Shaoting Guo, Wenqing Zhou, Yingying Luo, Yingqi Zhu, Lin Zhang, and Lei Li. Brain-model neural similarity reveals abstractive summarization performance. *Scientific Reports*, 15(1):370, January 2025. ISSN 2045-2322. doi: 10.1038/s41598-024-84530-w. URL https://www.nature.com/articles/s41598-024-84530-w. Publisher: Nature Publishing Group.

Riley Zurrin, Samantha Tze Sum Wong, Meighen M. Roes, Chantal M. Percival, Abhijit Chinchani, Leo Arreaza, Mavis Kusi, Ava Momeni, Maiya Rasheed, Zhaoyi Mo, Vina M. Goghari, and Todd S. Woodward. Functional brain networks involved in the raven's standard progressive matrices task and their relation to theories of fluid intelligence. *Intelligence*, 103:101807, March 2024. ISSN 0160-2896. doi: 10.1016/j.intell.2024.101807. URL https://www.sciencedirect.com/science/article/pii/S0160289624000011.

A Appendix

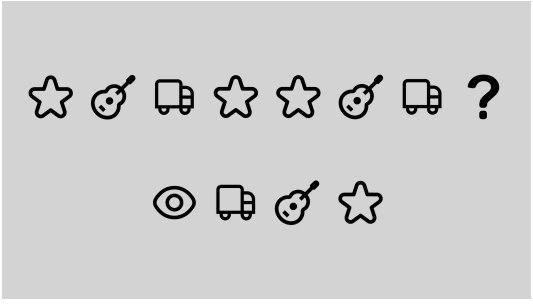


Figure 5: Example of an experimental trial for pattern type ABCAABCA. Top row: the sequence to be completed. Bottom row: four response options. The correct answer is the star icon (fourth position). Participants responded by pressing one of four computer keyboard buttons corresponding to each response option. FRPs were calculated when a new fixation was made to each icon in the sequence.

Full ID	Simplified ID	Parameter Count (billions)
microsoft/phi-4	Phi-4	14
google/gemma-2-2b-it	Gemma-2-2B	2
google/gemma-2-9b-it	Gemma-2-9B	9
google/gemma-2-27b-it	Gemma-2-27B	27
meta-llama/Llama-3.2-3B-Instruct	Llama-3.2-3B	3
meta-llama/Llama-3.3-70B-Instruct	Llama-3.3-70B	70
Qwen/Qwen2.5-72B-Instruct	Qwen2.5-72B	72
deepseek-ai/DeepSeek-R1-Distill-Llama-70B	Deepseek-R1-Distill-Llama-70B	70

Table 3: Open-source LLMs evaluated in this study. The "Full ID" column gives the full identifier used in the Hugging Face repository, while the "Simplified ID" correspond to the shortened version by which we refer to each LLM throughout this paper.

LLM PROMPT

A one-shot prompting strategy was employed, where an unrelated example and its associated answer were given to the LLM before showing it the actual problem to solve. Below is an example of a prompt used to query the LLMs:

> You will be presented with a sequence of words that follow a logical order, but one word in this sequence is missing, indicated by a question mark (?). Your task is to identify the missing word from a given set of options. To successfully complete this task, you should:
>
> 1. Analyze the sequence to understand the underlying logical pattern. Pay attention to the order of words, any repetitions, and how each word relates to the others in the sequence.
> 2. Do not rely on external tools or databases to analyze the sequence. Your reasoning should be based solely on the internal logic of the sequence as presented.
> 3. Consider the provided options carefully. There is only one correct answer that fits logically into the sequence in place of the question mark.
> 4. Present your answer in a clear and concise format: 'Answer: [chosen word]' (without the brackets). Include only your final choice in your response, without any additional explanation or text.
>
> Your goal is to determine which option logically completes the sequence. Remember, the key to solving this puzzle is understanding the pattern that links the words in the sequence. Use this pattern to decide which of the options fits as the missing word.
>
> Here is an example:
> Sequence: smile eye smile eye smile ?
> Options: camera eye bone smile
> Answer: eye
>
> Here is the puzzle you must now solve:
> Sequence: star star star guitar star star star ?
> Options: truck star cube guitar

EEG PREPROCESSING

EEG preprocessing was performed using the MNE-Python library (Gramfort et al., 2013) with the following steps:

1. Interpolation of manually identified bad electrodes
   Before any preprocessing, certain noisy or unreliable electrodes were manually flagged, either during data acquisiton or later inspection of the raw data. Instead of discarding those electrodes (which would create holes in the scalp map) we replaced them with a weighted average of their nearest neighbours using a spherical-spline algorithm (Perrin et al., 1989). This preserves the overall topography while preventing a handful of faulty sensors from biasing later steps.
2. Average re-referencing
   EEG measures voltage differences; choosing a reference that is itself noisy can contaminate the data of every electrode. By using the average activity of all electrodes as a reference

point we obtain a "neutral" zero-mean reference that minimises bias toward any single scalp location and improves the spatial interpretability of later analyses.

3. Notch filter (50–250 Hz in 50 Hz steps)
   The European electricity grid (Continental Europe Synchronous Area) operates at a standard frequency of 50 Hz. This creates noise at 50 Hz and multiples of it (harmonics) that gets mixed in with the EEG signal. We therefore used a series of narrow, "notch" filters centred on 50-Hz increments (from 50 to 250 Hz) to target and remove this noise while leaving neighbouring frequencies untouched. Each notch was implemented with MNE's default zero-phase, overlap-add finite impulse response (FIR) routine. Notch width followed the package default of frequency / 200 (e.g., 0.25 Hz at 50 Hz) with a fixed 1 Hz transition band. Filter length was chosen automatically (6.6 divided by the shortest transition bandwidth, using the default Hamming window), and edges were handled with MNE's "reflect_limited" padding, which mirrors the signal at each edge and then pads with zeros if necessary.
4. Independent component analysis (ICA) for artifact removal
   To isolate non-neural sources such as eye blinks, saccades, or muscle movements, we duplicated the data, applied a broad 1–100 Hz band-pass filter (which helps ICA converge to meaningful artifactual solutions), and ran an extended Infomax ICA (Langlois et al., 2010). This band-pass filter was implemented with the same FIR design choices described for the notch filters in Step 3 (zero-phase, overlap-add FIR; automatic kernel length, Hamming window, 'reflect_limited" padding). The resulting components were automatically labelled with the MNE-ICALabel classifier (Li et al., 2022). Components representing non-neural artefacts (i.e., muscle artifact, eye blink, heart beat, line noise, channel noise) were zeroed-out, and the cleaned mixing matrix was projected back onto the original unfiltered data, preserving genuine brain activity.
5. Band-pass filter (0.1–100 Hz)
   After artifact removal, a final band-pass filter from 0.1 to 100 Hz was applied to the data. This step ensured the removal of any remaining low-frequency drift and high-frequency noise while preserving the physiologically relevant neural oscillations. This band-pass filter was, again, implemented with the same FIR design choices described for the notch filters in Step 3.
6. Final average re-referencing
   To ensure optimal signal quality and consistency across all electrodes, the data underwent a second average re-referencing step. This final re-referencing minimized any residual common mode noise that might have been introduced or remained after previous processing steps, preparing the data for subsequent analyses.

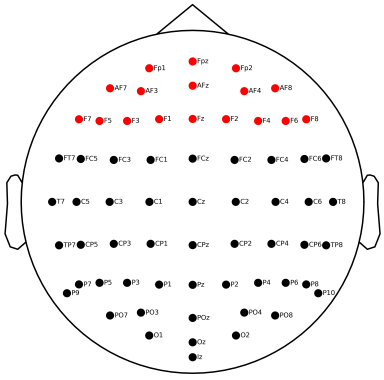


Figure 6: EEG montage used in this experiment (international 10-20 system; 64-electrodes). Frontal electrodes used in the analysis are highlighted in red.

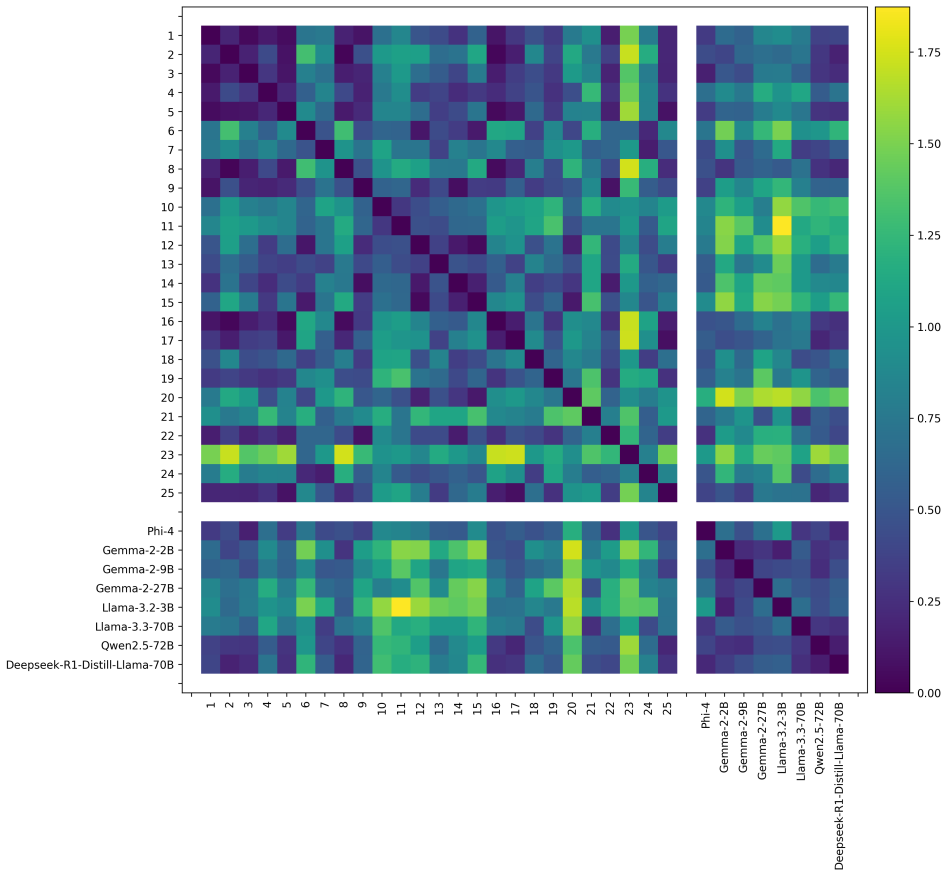


Figure 7: Accuracy RDM, built from the average accuracy by pattern type for each human (labels 1-25) and LLM; brighter colors denote greater dissimilarity.

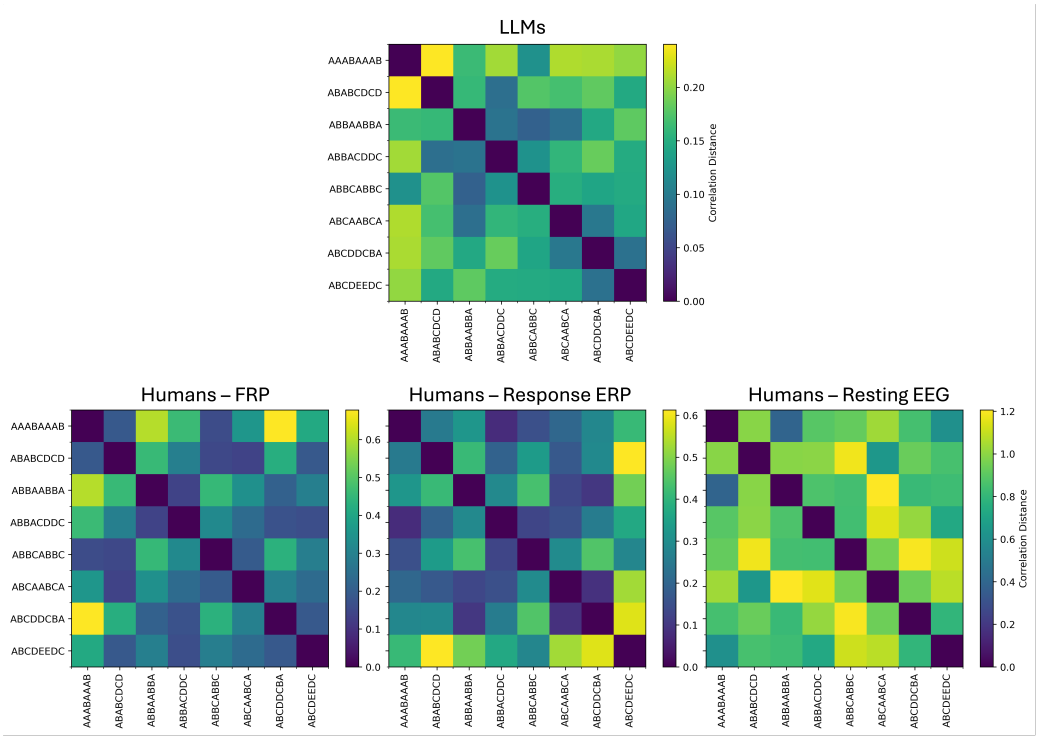


Figure 8: Pattern-level RDMs. Top: group-average RDM from LLMs; Bottom: group-average RDM from human participants.

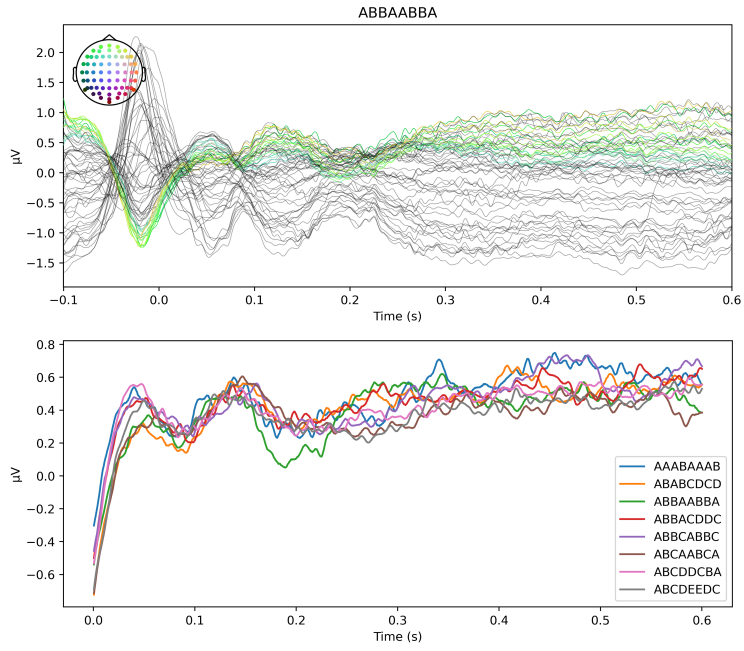


Figure 9: Frontal FRPs during the "decision phase" (see Task design section in the Methods). Top: Example of a single-pattern grand average FRP (N = 25 participants) elicited by fixations on trials of the ABBAABBA pattern type. Each trace corresponds to one EEG channel; frontal electrodes are shown in color, all other channels in grey. The inset scalp map (nose up) at the top left displays the electrodes spatial arrangement on the scalp. Time 0 s marks fixation onset. Bottom: Pattern-specific frontal traces. FRPs averaged across all participants, trials, and frontal electrodes for each of the eight abstract patterns.