Large Language Models Show Signs of Alignment with Human Neurocognition During Abstract Reasoning

Anonymous Submission

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.

Keywords: Abstract Reasoning; AI; Artificial Neural Networks; NeuroAI; Deep Learning; Large Language Models (LLMs); Electroencephalography (EEG); Fixation Related Potentials (FRPs); Representational Similarity Analysis (RSA)

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 and 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). However, if LLMs' reasoning is truly human-like, we should not only expect human-like responses but also human-like internal representations and computational processes. To investigate this, we extend recent work on brain–model alignment 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 and 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 and 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 and 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 and Mitchell, 2024; Li et al., 2025; Liang et al., 2025; McCoy et al., 2024; Mitchell et al., 2023; Nguyen et al., 2025; Palmarini and Mitchell, 2024; Sourati et al., 2024; Stevenson et al., 2025; Yang et al., 2025; Yax et al., 2024). Even when correct, models often rely on brittle heuristics rather than rule manipulation (Lewis and Mitchell, 2024; Lee et al., 2025; Chollet, 2019).

Most evaluations focus on outcomes, leaving open whether LLMs develop human-like internal abstractions. Addressing this requires comparing their representational structure to brain activity during reasoning. 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 and Liversedge, 2020). Their rapid frequency during free viewing also allows high data yield in short sessions.

Given the established role of frontal regions in abstract reasoning, we focused EEG analysis on frontal electrodes. FRPs from these sites 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., ABCDDCBA). 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.

To our knowledge, this is the first study to test whether LLMs and the human brain converge on shared representational geometries during abstract reasoning.

Figure 1. 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.

RESULTS

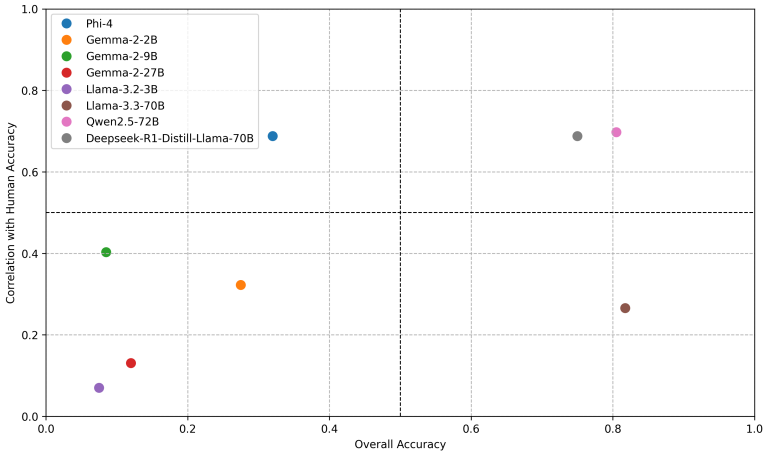

Figure 2. LLMs' overall task accuracy vs. correlation with human accuracy profile. Each dot represents one LLM and is positioned by its overall accuracy on the 400 abstract-sequence trials (x-axis) and the Pearson correlation (r) between its 8-pattern accuracy vector and that of human participants (y-axis).

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 2). 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 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.

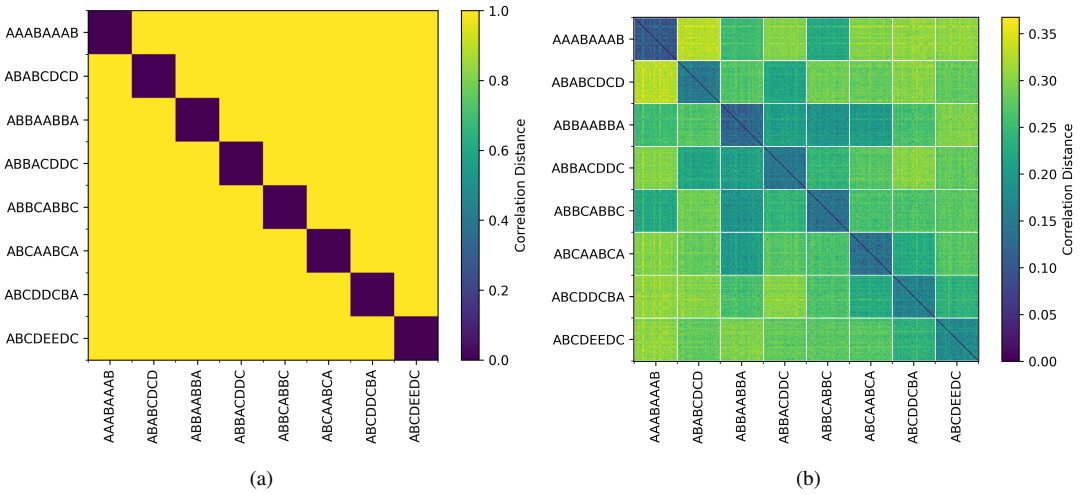

Figure 3. Trial-level representational dissimilarity matrices (RDMs) generated from 400 unique trials (50 trials per pattern type). **Left:** Task-optimal Reference RDM, encoding perfect within-pattern similarity and maximal between-pattern dissimilarity; **Right**: grand-average LLM RDM, computed from the activations of each LLM's task-optimal layer; brighter colors denote greater dissimilarity.

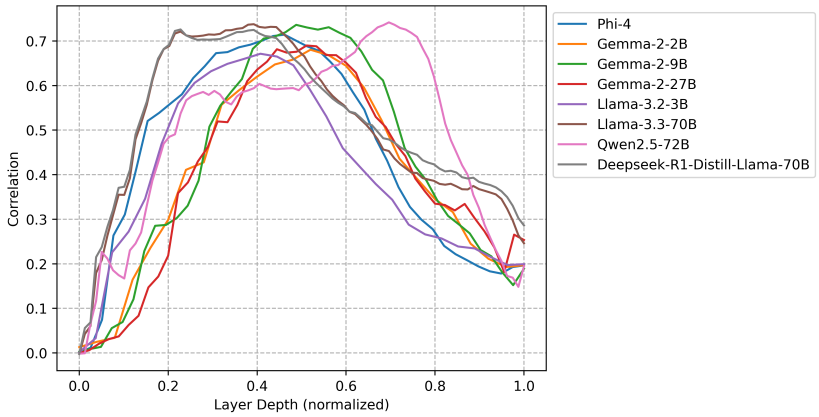

Figure 4. Layer-wise Pearson correlation between LLM layer RDMs and reference RDM.

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 3(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.

We found that layer-wise correlations to this reference RDM form a pronounced inverted-U profile for each LLM (Figure 4): 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 3(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.

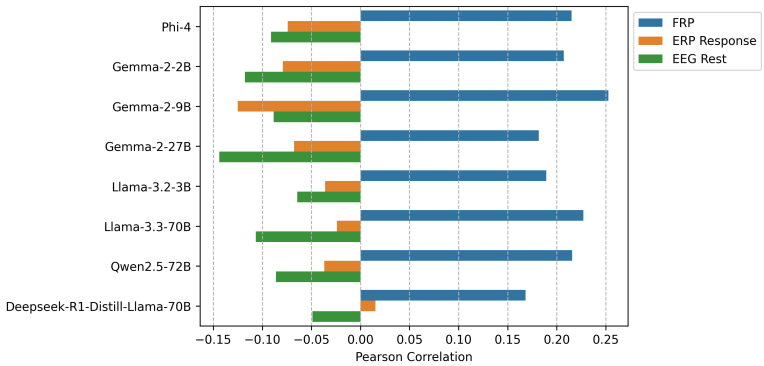

Figure 5. Pattern-level RSA between Human group-averaged RDMs and LLMs' task-optimal layer RDMs

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 5 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).

	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

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

Resting ERPs were consistently negative, r ≈ –.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 ≈ .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 makes a model succeed on the task are the same representations that seem to faintly reappears in human frontal FRPs, hinting at a shared coding of abstract patterns.

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 4). 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 and Cooper, 2025; Meng et al., 2023; Park and 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 3(b) and Figure 7), which is also corroborated by earlier findings (Lan et al., 2025; Wolfram and 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 and 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.

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: three withdrew after their first session, one after two sessions, and another after three sessions. 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 1), disregarding those that landed on the choice options (bottom row of Figure 1).

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 $\times$ 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 3(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 $\times$ 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 representational similarity analysis (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.

REFERENCES

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., and Blything, R. (2022). Deep Problems with Neural Network Models of Human Vision. *The Behavioral and Brain Sciences*, pages 1–74.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of artificial general intelligence: early experiments with GPT-4. arXiv:2303.12712 [cs].

Caucheteux, C. and King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):1–10. Publisher: Nature Publishing Group.

Caudle, M. M., Spadoni, A. D., Schiehser, D. M., Simmons, A. N., and Bomyea, J. (2023). Neural activity and network analysis for understanding reasoning using the matrix reasoning task. *Cognitive Processing*, 24(4):585–594.

Choi, Y. Y., Shamosh, N. A., Cho, S. H., DeYoung, C. G., Lee, M. J., Lee, J.-M., Kim, S. I., Cho, Z.-H., Kim, K., Gray, J. R., and Lee, K. H. (2008). Multiple bases of human intelligence revealed by cortical thickness and neural activation. *Journal of Neuroscience*, 28(41):10323–10329.

Chollet, F. (2019). On the measure of intelligence. arXiv:1911.01547 [cs].

Chuderski, A. (2022). Fluid intelligence emerges from representing relations. *Journal of Intelligence*, 10(3):51.

Degno, F. and Liversedge, S. P. (2020). Eye movements and fixation-related potentials in reading: a review. *Vision*, 4(1):11.

Dima, D. C., Janarthanan, S., Culham, J. C., and Mohsenzadeh, Y. (2024). Shared representations of human actions across vision and language. *Neuropsychologia*, 202:108962.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., and Charest, I. (2024). Visual representations in the human brain are aligned with large language models. arXiv:2209.11737 version: 2.

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, 14(4):172–179.

Engbert, R. (2006). Microsaccades: a microcosm for research on oculomotor control, attention, and visual perception. In Martinez-Conde, S., Macknik, S. L., Martinez, L. M., Alonso, J. M., and Tse, P. U., editors, *Progress in Brain Research*, volume 154 of *Visual Perception*, pages 177–192. Elsevier.

Feghhi, E., Hadidi, N., Song, B., Blank, I. A., and Kao, J. C. (2024). What Are Large Language Models Mapping to in the Brain? A Case Against Over-Reliance on Brain Scores. arXiv:2406.01538 [cs] version: 1.

Ferrer, E., O'Hare, E. D., and Bunge, S. A. (2009). Fluid reasoning and the developing brain. *Frontiers in Neuroscience*, 3(1):46–51.

Gawin, C., Sun, Y., and Kejriwal, M. (2025). Navigating semantic relations: challenges for language models in abstract common-sense reasoning. arXiv:2502.14086 [cs].

Gendron, G., Bao, Q., Witbrock, M., and Dobbie, G. (2024). Large language models are not strong abstract reasoners. arXiv:2305.19555 [cs].

Gluth, S., Rieskamp, J., and Büchel, C. (2013). Classic EEG motor potentials track the emergence of value-based decisions. *Neuroimage*, 79:394–403.

Goldstein, A., Grinstein-Dabush, A., Schain, M., Wang, H., Hong, Z., Aubrey, B., Schain, M., Nastase, S. A., Zada, Z., Ham, E., Feder, A., Gazula, H., Buchnik, E., Doyle, W., Devore, S., Dugan, P., Reichart, R., Friedman, D., Brenner, M., Hassidim, A., Devinsky, O., Flinker, A., and Hasson, U. (2024). Alignment of brain embeddings and artificial contextual embeddings in natural language points to common geometric patterns. *Nature Communications*, 15(1):2768. Publisher: Nature Publishing Group.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-python. *Frontiers in Neuroscience*, 7:267.

Gray, J. R., Chabris, C. F., and Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, 6(3):316–322. Publisher: Nature Publishing Group.

Hersche, M., Camposampiero, G., Wattenhofer, R., Sebastian, A., and Rahimi, A. (2024). Towards learning to reason: comparing LLMs with neuro-symbolic on arithmetic relations in abstract reasoning. arXiv:2412.05586 [cs] version: 1.

Huh, M., Cheung, B., Wang, T., and Isola, P. (2024). The platonic representation hypothesis. arXiv:2405.07987 [cs].

Iaia, C., Choksi, B., Wiebers, E., Roig, G., and Fiebach, C. J. (2025). The representational alignment between humans and language models is implicitly driven by a concreteness effect. arXiv:2505.15682 [cs] version: 1.

Ju, T., Sun, W., Du, W., Yuan, X., Ren, Z., and Liu, G. (2024). How large language models encode context knowledge? A layer-wise probing study. arXiv:2402.16061 [cs].

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2. Publisher: Frontiers.

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

Lan, M., Torr, P., Meek, A., Khakzar, A., Krueger, D., and Barez, F. (2025). Quantifying feature space universality across large language models via sparse autoencoders. arXiv:2410.06981 [cs] version: 4.

Langlois, D., Chartier, S., and Gosselin, D. (2010). An introduction to independent component analysis: InfoMax and FastICA algorithms. *Tutorials in Quantitative Methods for Psychology*, 6(1):31–38.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. Publisher: Nature Publishing Group.

Lee, S., Sim, W., Shin, D., Seo, W., Park, J., Lee, S., Hwang, S., Kim, S., and Kim, S. (2025). Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. *ACM Trans. Intell. Syst. Technol.*, page 3712701. Just Accepted.

Lei, G. and Cooper, S. J. (2025). The representation and recall of interwoven structured knowledge in LLMs: a geometric and layered analysis. arXiv:2502.10871 [cs] version: 1.

Lei, Y., Ge, X., Zhang, Y., Yang, Y., and Ma, B. (2025). Do large language models think like the brain? Sentence-level evidence from fMRI and hierarchical embeddings. arXiv:2505.22563 [cs].

Lewis, M. and Mitchell, M. (2024). Evaluating the robustness of analogical reasoning in large language models. arXiv:2411.14215 [cs].

Li, A., Feitelberg, J., Saini, A. P., Höchenberger, R., and Scheltienne, M. (2022). MNE-ICALabel: automatically annotating ICA components with ICLabel in python. *Journal of Open Source Software*, 7(76):4484.

Li, Y., Gao, Q., Zhao, T., Wang, B., Sun, H., Lyu, H., Hawkins, R. D., Vasconcelos, N., Golan, T., Luo, D., and Deng, H. (2025). Core knowledge deficits in multi-modal language models. arXiv:2410.10855 [cs].

Liang, S., Garg, S., and Moghaddam, R. Z. (2025). The SWE-bench illusion: when state-of-the-art LLMs remember instead of reason. arXiv:2506.12286 [cs].

Lui, K. K., Nunez, M. D., Cassidy, J. M., Vandekerckhove, J., Cramer, S. C., and Srinivasan, R. (2021). Timing of readiness potentials reflect a decision-making process in the human brain. *Computational Brain & Behavior*, 4(3):264–283.

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., and Griffiths, T. L. (2024). Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445. Publisher: Nature Publishing Group.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., and Griffiths, T. L. (2024). 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. Publisher: Proceedings of the National Academy of Sciences.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2023). Locating and editing factual associations in GPT. arXiv:2202.05262 [cs].

Mischler, G., Li, Y. A., Bickel, S., Mehta, A. D., and Mesgarani, N. (2024). Contextual Feature Extraction Hierarchies Converge in Large Language Models and the Brain.

Mitchell, M., Palmarini, A. B., and Moskvichev, A. (2023). Comparing humans, GPT-4, and GPT-4V on abstraction and reasoning tasks. arXiv:2311.09247 [cs].

Musker, S., Duchnowski, A., Millière, R., and Pavlick, E. (2025). LLMs as models for analogical reasoning. arXiv:2406.13803 [cs] version: 2.

Newell, A. (1955). 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), pages 101–108, New York, NY, USA. Association for Computing Machinery.

Newell, A. and Simon, H. (1956). The logic theory machine–a complex information processing system. *IRE Transactions on Information Theory*, 2(3):61–79.

Nguyen, T. D., Watts, D. J., and Whiting, M. E. (2025). Empirically evaluating commonsense intelligence in large language models with large-scale human judgments. arXiv:2505.10309 [cs].

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553.

Palmarini, A. B. and Mitchell, M. (2024). Abstract understanding of core-knowledge concepts: Humans vs. llms. In *ICML 2024 Workshop on LLMs and Cognition*.

Park, H. and Kim, G. (2025). Where do LLMs encode the knowledge to assess the ambiguity? In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., Schockaert, S., Darwish, K., and Agarwal, A., editors, *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 445–452, Abu Dhabi, UAE. Association for Computational Linguistics.

Perfetti, B., Saggino, A., Ferretti, A., Caulo, M., Romani, G. L., and Onofrj, M. (2007). Differential patterns of cortical activation as a function of fluid reasoning complexity. *Human Brain Mapping*, 30(2):497–510.

Perrin, F., Pernier, J., Bertrand, O., and Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, 72(2):184–187.

Santarnecchi, E., Emmendorfer, A., and Pascual-Leone, A. (2017). Dissecting the parieto-frontal correlates of fluid intelligence: a comprehensive ALE meta-analysis study. *Intelligence*, 63:9–28.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. (2021). 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.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489. Publisher: Nature Publishing Group.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144. Publisher: American Association for the Advancement of Science.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359. Publisher: Nature Publishing Group.

Skean, O., Arefin, M. R., LeCun, Y., and Shwartz-Ziv, R. (2024). Does representation matter? Exploring intermediate layers in large language models. arXiv:2412.09563 [cs].

Skean, O., Arefin, M. R., Zhao, D., Patel, N., Naghiyev, J., LeCun, Y., and Shwartz-Ziv, R. (2025). Layer by layer: uncovering hidden representations in language models. arXiv:2502.02013 [cs].

Sourati, Z., Ilievski, F., Sommerauer, P., and Jiang, Y. (2024). ARN: analogical reasoning on narratives. arXiv:2310.00996 [cs].

Stevenson, C. E., Pafford, A., Maas, H. L. J. v. d., and Mitchell, M. (2025). Can large language models generalize analogy solving like people can? arXiv:2411.02348 [cs].

Sun, W., Song, X., Li, P., Yin, L., Zheng, Y., and Liu, S. (2025). The curse of depth in large language models. arXiv:2502.05795 [cs].

Tschentscher, N., Mitchell, D., and Duncan, J. (2017). Fluid intelligence predicts novel rule implementation in a distributed frontoparietal control network. *Journal of Neuroscience*, 37(18):4841–4847.

Wang, Y., Chen, W., Han, X., Lin, X., Zhao, H., Liu, Y., Zhai, B., Yuan, J., You, Q., and Yang, H. (2024). Exploring the reasoning abilities of multimodal large language models (MLLMs): a comprehensive survey on emerging trends in multimodal reasoning. arXiv:2401.06805 [cs].

Webb, T., Holyoak, K. J., and Lu, H. (2023). Emergent Analogical Reasoning in Large Language Models.

Webb, T. W., Holyoak, K. J., and Lu, H. (2025). Evidence from counterfactual tasks supports emergent analogical reasoning in large language models. *PNAS Nexus*, 4(5):pgaf135.

Williams, S. and Huckle, J. (2024). Easy problems that LLMs get wrong. arXiv:2405.19616 [cs].

Wolfram, C. and Schein, A. (2025). Layers at similar depths generate similar activations across LLM architectures. arXiv:2504.08775 [cs].

Yang, Y., Chen, M., Liu, Q., Hu, M., Chen, Q., Zhang, G., Hu, S., Zhai, G., Qiao, Y., Wang, Y., Shao, W., and Luo, P. (2025). Truly assessing fluid intelligence of large language models through dynamic reasoning evaluation. arXiv:2506.02648 [cs] version: 1.

Yax, N., Anlló, H., and Palminteri, S. (2024). Studying and improving reasoning in humans and machines.

Communications Psychology, 2(1):51. Publisher: Nature Publishing Group.

Zhang, Y., Dong, Y., and Kawaguchi, K. (2024). Investigating layer importance in large language models. In *Proceedings of the 7th Blackboxnlp Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 469–479, Miami, Florida, US. Association for Computational Linguistics.

Zhang, Z., Guo, S., Zhou, W., Luo, Y., Zhu, Y., Zhang, L., and Li, L. (2025). Brain-model neural similarity reveals abstractive summarization performance. *Scientific Reports*, 15(1):370. Publisher: Nature Publishing Group.

Zurrin, R., Wong, S. T. S., Roes, M. M., Percival, C. M., Chinchani, A., Arreaza, L., Kusi, M., Momeni, A., Rasheed, M., Mo, Z., Goghari, V. M., and Woodward, T. S. (2024). Functional brain networks involved in the raven's standard progressive matrices task and their relation to theories of fluid intelligence. *Intelligence*, 103:101807.