Zero-shot neural predictivity in human prefrontal cortex with a massively multi-task multimodal transformer

Lucas Gomez
Integrated Program in Neuroscience
McGill University & Mila
lucas.gomez@mail.mcgill.ca
https://www.lucasgomez.ca/

Aziz Ktari
Department of Psychology
McGill University
aziz.ktari@mail.mcgill.ca
https://hancilar.github.io/Aziz.ktari

Hao Yuan Bai
Research Intern
McGill University & Mila
hao.bai@mcgill.ca

Pouya Bashivan
Department of Physiology
McGill University & Mila
pouya.bashivan@mcgill.ca

Abstract

Working memory supports a broad spectrum of behaviors and higher cognitive abilities, with the prefrontal cortex playing a central role in this capacity. Although prior work has identified which brain regions are engaged in specific working memory tasks, and in some cases how they contribute, we still lack a general framework that can predict which regions will be recruited in novel tasks, what information they represent, and the computations they perform. To address this gap, we trained a single neural network on millions of visual decision-making tasks with sensory-realistic inputs, aiming to build a generalized model of working memory. We evaluated the model against an fMRI dataset spanning 12 tasks and hundreds of distinct conditions, testing its ability to capture neural activity across the brain, with a focus on the prefrontal cortex. Our results show that large models trained on a broad distribution of tasks can predict brain activity zero-shot, outperforming even models trained directly on the target tasks. This ability improves further with model size, which consistently enhances prediction accuracy. Furthermore, analyses of layer-to-region correspondences largely conformed with the theories of hierarchical organization along the rostro-caudal axis of the prefrontal cortex. These findings suggest that neural network models hold significant potential not only for simulating neural activity in regions previously difficult to model, but also for revealing how the brain encodes, organizes, and manipulates task information during working memory.

1 Introduction

Working memory (WM) is a key component of human cognition, allowing us to temporarily store and manipulate task-relevant information. While WM engages a wide network of brain regions, the prefrontal cortex (PFC) plays a central role. PFC is responsible for the executive function and has been implicated in a wide variety of goal-directed behaviours that include selective processing of information (i.e. attention) and making choices between alternative choices. PFC disruptions are linked to conditions such as schizophrenia, ADHD, depression, and cognitive decline.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Data on the Brain & Mind.

Thanks to neuroimaging and electrophysiology experiments, much is known about the role of different nodes in the working memory brain network in a variety of different cognitive tasks. However, despite these advances, the current accounts only broadly capture the high-level description of the overall functions and our understanding of the underlying mechanisms that give rise to these representations remain largely unspecified.

For instance, we know that much like many other brain regions, specially in sensory cortices [19, 23], the PFC is thought to be hierarchically organized along the rostro-caudal axis of the frontal lobe [5, 22]. In this framework, more rostral regions are proposed to represent increasingly abstract and complex rules, domain general memories, as well as actions and decisions that unfold over longer timescales. A range of neuroimaging studies have found evidence supporting this hierarchical organization of the PFC [6, 8, 29, 30]. Other work has further described the functional selectivity across PFC sub-regions [3, 7, 12, 25, 32]. Nevertheless, debate remains regarding the uniformity of this abstraction gradient and whether a precise, general account of functional selectivity across PFC sub-regions that holds consistently across cognitive tasks can be established.

The general gap in understanding of the PFC compared to other cortical areas may stem from several factors, including the broad and abstract nature of the information the PFC encodes during cognitive tasks, as well as the current reliance on purely descriptive models and the absence of computational frameworks capable of capturing WM across diverse task settings.

Over the past decade, work at the intersection of neuroscience and AI has shown that artificial neural networks (ANNs) trained on naturalistic tasks can develop internal activations aligned with brain activity [20, 24, 41]. Such models not only mirror neural dynamics but also offer new ways to predict and modulate neural activity [9, 34, 39]. However, *there are currently no generalizable ANNs that can model WM and decision-making in the PFC*, with the existing models having limited utility and scope as a result of narrowly defined tasks and use of symbolic inputs [10, 33, 43].

To address the generalization gap in existing models of working memory and decision making, we developed a neural network model capable of solving a wide variety of visual decision making tasks from natural images and text. To do this, we trained a large multimodal neural network model on millions of visual decision-making tasks with natural language instructions.

To assess the model's alignment with the brain, we measured how well the internal activity of units in the model can predict neural activity in various brain regions and compared it against a range of other baseline models using newly collected whole-brain human fMRI WM dataset [2]. Using the best model's layer-to-region mappings, we investigate the prominent theory of a rostral-caudal axis of hierarchical organization within the PFC.

Our contributions are as follows:

1. We developed a large multi-task multi-modal neural network model trained on millions of visual working memory tasks, demonstrating strong generalization to unseen tasks.
2. We evaluated neural similarity between our model and the brain at both the whole-brain network and granular PFC levels, benchmarking it against a range of baselines.
3. By examining the model's internal representations and their correspondence to different PFC sub-regions, we explored the representational hierarchy along the rostral-caudal axis of the PFC.

2 Related works

Many models of working memory have been proposed across cognitive science and neuroscience. Early accounts primarily provided conceptual frameworks describing how the brain might support working memory. Among the most influential are Baddeley's model [4] and Cowan's model [16]. These were not computationally implementable, however, later work introduced frameworks with tunable components (e.g., computer code) that could be hand-designed to perform specific tasks [27, 35]. More sophisticated approaches employed neural network models that used optimization techniques, to learn from experience, by adjusting their parameters without manual tuning [31].

Artificial neural networks, particularly recurrent neural networks, have since been widely applied to simulate working memory across different scales. These studies demonstrated that recurrent dynamics can reproduce persistent activity, a hallmark of working memory [13, 15]. Yet these models were typically limited to one or a few cognitive tasks. Subsequent work showed that the same class of models could be trained to solve dozens of tasks, highlighting how their internal dynamics enable the reuse of similar motifs across contexts [18, 42, 44].

Despite these advances, computational models that can handle realistic inputs (e.g., natural images and text), while maintaining generality across many tasks, remain rare. A notable recent effort, *Centaur* [11], was shown to replicate human behavior across a wide range of decision-making tasks. However, its ability to generalize to unseen tasks and individuals was limited [28], and the model was restricted to text-based paradigms. In parallel, vision-language models have achieved impressive progress on problems requiring both visual and textual inputs, but many still fail at seemingly simple tasks that demand attending to specific object properties and selectively recalling them based on task instructions [26, 40].

3 Methods

3.1 Data

3.1.1 Working memory task data

To train the neural network model, we required a large and diverse dataset of tasks. We specifically desired the tasks to involve natural visual stimuli, natural language task instructions, selective attention to particular object features, and diverse reasoning requirements. To satisfy these demands we used iWISDM [26], a recently developed environment capable of generating a virtually unlimited number of visual decision-making tasks, to generate all the data for training and validating our models. iWISDM uses a compositional structure to construct tasks using various operators and combining simple task blocks in various ways to build increasingly more complex ones.

For the large-scale model training datasets we use iWISDM's `AutoTask` mode, which allows for the controlled generation of millions of random tasks varying in complexity. Using iWISDM, each task can be rendered into a trial provided as a series of images, corresponding task instructions in natural language, and ground-truth answers (see Appendix Figure 10 for iWISDM task examples). For our purposes, a task trial consists of three elements: (1) a series of frames (max 10), with each frame being blank (delay) or containing a single everyday object (e.g., bus, bench, chair, couch) placed in one of four locations; (2) The trial instructions describe the working memory-based decision-making task in natural language; (3) A sequence of actions that specify the correct outputs at every step. The visual objects shown on the frames are 3D objects selected from the Shapenet-3D dataset [14], and are rendered at varying viewing angles. The rendering of faces and buses shown to participants from the fMRI dataset are also added to this stimulus set. The large-scale model is trained on a total of 80,730 base-tasks. However, during training, as these base-tasks are randomly selected and merged to create a single trial, the model sees millions of unique tasks. A set of 11,470 held-out base-tasks is used for validation.

3.1.2 Working memory fMRI data

For our neural analyses, we used a recently collected yet to be published functional MRI (fMRI) dataset that contains whole-brain scans of 4 human subjects performing 12 working memory tasks [2]. The data was collected with a Siemens Prisma 3T scanner and contains ~ 8 hours of scan time per subject. The data consists of recordings from each subject performing 2128 trials from the 12 tasks. The task trials are rendered with stimuli consisting of two faces and two buses at two unique viewing angles, see Appendix Figure 11 for examples of the tasks and stimuli.

Importantly, every fMRI trial consists of encoding and delay frames. An encoding frame is any image that shows an object needed to solve the task. A delay frame is the blank image that immediately follows an encoding frame. Voxel activity was binned into time points aligned to when (a) an encoding frame was shown or (b) a delay frame was shown.

To reduce noise and improve reliability, trials were grouped into task-conditions defined jointly by task structure, stimuli pairs, and stimuli positions within the task sequence. For example, the (Bus1-Face1) pair presented as the first two stimuli in the contextual decision-making task was treated as a unique condition. Averaging over repeated trials, within each task-condition, substantially enhanced the signal-to-noise ratio, yielding 416 distinct task-conditions for neural prediction.

3.2 Model

3.2.1 Architecture

Our proposed model is a neural network model consisting of three primary components:

1. **Image Encoder**: This component receives the trial images and encodes them into image embedding tokens. In both the large-scale and baseline experiments the activations from the last non-pooling layer of a frozen EfficientNetV2-M [38] are used after being convolved.
2. **Instruction Encoder**: Takes in the trial text instructions and encodes them as a series of language embedding tokens using a frozen jina-embeddings-v3 model [37].
3. **Core Action Decoder**: Takes in the encoded trial images and text instructions as inputs and outputs the ground-truth outputs to the task at each step. The large-scale model uses a transformer for this module, however any sequence2sequence model is viable.

The model is defined as

$$g(\mathbf{i}, \mathbf{j}) = [L_1(E_{ins}(\mathbf{i})),\ Conv(E_{img}(\mathbf{f}))], \quad f(g(\mathbf{i}, \mathbf{j})) = L_2(D_{act}(g(\mathbf{i}, \mathbf{j}))),$$

where $\mathbf{i}$ is a text instruction (encoded as T tokens via E_{ins}), $\mathbf{f}$ a sequence of N images encoded via E_{img}, $L_1 : (T, d_{ins}) \rightarrow (T, d_{hidden})$ is a linear projection, $Conv : (N, d_{img}, H, W) \rightarrow (N, d_{hidden}, 1, 1)$ a 3D-convolutional projection, and $L_2 : (N, d_{hidden}) \rightarrow (N, \text{num\_actions})$ another linear projection taking as input the decoded tokens from D_{act}. Here g outputs the concatenated sequence of encoded instruction and image tokens, and f outputs the action class logits for each image in the trial.

The large-scale model contains a total of 482,831,363 million trainable parameters all originating in the Core Action Decoder. For a detailed report of the large-scale and baseline architectures see Appendix Table 1.

3.2.2 Baseline architecture

The models used as baselines to compare with the generalized large-scale transformer are nearly identical with exception of two modifications: (1) we replace the text-instructions with a one-hot task-vector which encodes which of the 12-fMRI tasks the trial is from, and (2) LSTM [21] and Mamba2 [17] models are used instead of a transformer for the Core Action Decoder component in two of our baselines.

3.2.3 Training procedure

The large-scale model is trained for 1,500 epochs and an effective batch-size of 1,400, where each epoch defines a model sweep through 14,000 rendered task trials. On every epoch, these 14,000 trials are generated from tasks that are created by randomly merging a random number of the 80,730 training base-tasks. Cosine scheduling is used with a peak learning-rate of $3.5e^{-5}$ and 150 iterations of warmup. See Appendix Table 1 for more details on the large-scale and baseline model training regimes.

3.3 Neural prediction

To assess the correspondence between model representations and neural activity, we present all trials from the fMRI dataset to each model while recording their unit activity. For transformer architectures, we extract activations from the final MLP layer of each transformer block. For state-based models such as Mamba2 and LSTM, we use their hidden states.

For each trial, we construct encoding-delay pair representations by concatenating the image position tokens from consecutive encoding and delay periods along the model's hidden feature dimension. Using PCA ridge regression, these encoding-delay pair representations are used to predict beta values from the corresponding delay time steps. We excluded encoding periods from prediction analyses, as they showed no reliable predictive signal in our experiments.

For the analyses at network and PFC region-of-interest (ROI) levels, voxel-wise betas are averaged within the network or ROI respectively, and the voxel-averaged values are predicted as described for every subject. Prediction performance is quantified as the Pearson correlation between the predicted and observed bteas, computed in a 10-fold cross-validation scheme and averaged across subjects. This procedure results in r-value correlation scores from every model layer to every network or PFC ROI.

To account for measurement reliability, we estimate noise-ceilings (NC) for each subject using Spearman-Brown corrected split-half reliability tests [36]. Noise-ceilings are computed separately for each cortical network and PFC ROI, before being averaged across the four subjects. All model activation prediction correlations are corrected by dividing by these NC values.

3.4 PFC hierarchy analysis

We first evaluate each model layer with each PFC ROI using our linear mapping approach described above. We then pick the layer with the highest predictivity score for each ROI as the best layer explaining that ROI. To quantify the level of abstraction encoded in representations from each layer of the model, we define a two-level task abstraction hierarchy.

At the lower level (Level 1), task information is defined by object features (location and category). At the higher level (Level 2), task information is defined by task logic, such as the outcome of a subtask (e.g. 'location of object 2 equals location of object 1?').

To measure the degree of abstraction encoded at each model layer, we treat this as a multi-class prediction problem. Following a similar procedure to the one described for neural prediction, we use PCA ridge regression, but this time to predict both object-feature and task-logic information from the fMRI trials, using activations from each corresponding model layer. This allows us to assess how hierarchical task abstraction emerges across model depth and its correspondence with PFC ROIs.

We use the object-feature and task-logic decoding accuracies from each layer to derive a single metric that reflects where the representations lie along the abstraction hierarchy. Specifically, for each model layer l we compute a hierarchy score defined as $h_l = -a_{l,O} + a_{l,T}$, where $a_{l,O}$ is the average object-feature decoding accuracy and $a_{l,T}$ is the average task-logic decoding accuracy for layer l. The resulting h_l values range from -1.0 to 1.0, providing a continuous measure of hierarchical encoding: negative values indicate representations biased toward low-level object features, whereas positive values indicate representations biased toward higher-level task-logic information.

4 Results

4.1 Network level prediction

We assess the neural similarity of our main model relative to several baseline models by using their internal activations to predict average activity across canonical brain networks [1]. We find that the large, more generalized Transformer model (Big iWISDM TF) outperforms all baselines across every network (see Figure 1). The large model trained only on the 12 fMRI tasks (Big fMRI-Only) also consistently exceeds the smaller version, underscoring the joint importance of model size and training generality for achieving brain-like representations. The small Mamba2-based fMRI-only model (Small fMRI-Only Mamba) consistently outperforms the other small-scale baselines, despite having the fewest trainable parameters (3.8M), suggesting it is a promising architecture for capturing neural representations, at least at small-scales.

Figure 2 shows the best-layer NC corrected correlations of the large-scale generalized Transformer model mapped onto the cortical surface. Prediction accuracy is consistently higher in fronto-lateral, centro-lateral, and centro-medial regions compared to fronto-medial and posterior-medial regions. Supplementary Figure 8 shows the corresponding maps for the other trained models. Across all models, the relative distribution of representational similarity across regions remains consistent, with differences emerging primarily in the overall strength of alignment.

4.2 Prefrontal cortex region-level prediction

Figure 3 presents the highest final NC corrected correlation values across models and layers for 34 regions spanning the PFC. Compared to the network-level analyses, there is greater variability in prediction performance across individual PFC ROIs. Nonetheless, the average NC corrected

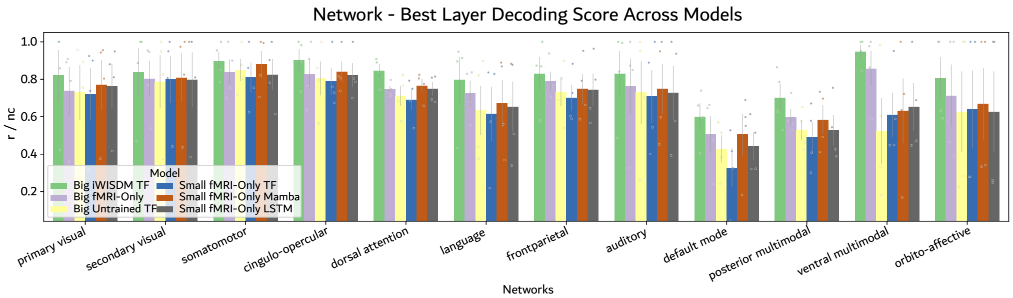


Figure 1: The best neural predictivity scores (NC corrected r-correlation) across model layers and networks for every model tested.

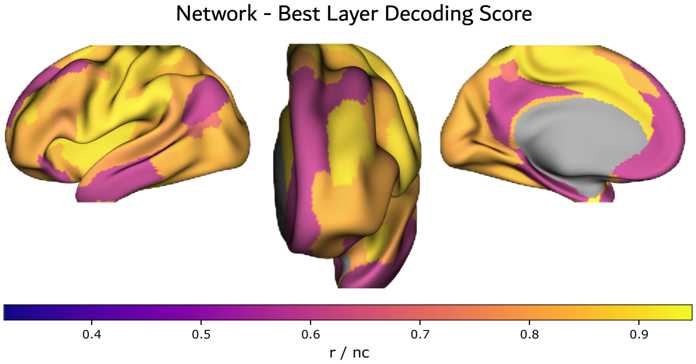


Figure 2: A network surface map of the best neural predictivity scores (NC corrected r-correlation) across layers of the Big iWISDM TF model.

prediction correlation across regions for the Big iWISDM Transformer is 0.79, with some regions reaching nearly 1.0 correlation. Notably, the less generalized version of the Big Transformer, trained only on the 12 fMRI tasks, approaches the performance of the best model in nearly half of the regions. This suggests that the neural similarity benefits of building broad, generalized representations diminish as the spatial granularity of the neural target increases. However, it is also important to note that the large modelis performing these 12-tasks zero-shot as opposed to it's fMRI-only counterparts.

Overall, these results reinforce the conclusion that both model size and the degree of task generalization remain key determinants of neural similarity within the PFC.

Looking closer we find that dorsolateral PFC activity is predicted especially well by model activations (see Figure 4). For example, Brodmann Area (BA) 46 known to play a central role in decision-making, attention, and working memory is predicted with a NC corrected correlation of 0.90. More broadly, these results maintain the pattern that models trained on ecological tasks yield representations closely aligned with the brain regions engaged by those tasks.

4.3 Prefrontal cortex hierarchy analyses

We find layers of the most generalized large transformer that best predict PFC ROIs are layers 1-4, with middle layers 2 and 3 serving as the best predictors for 65% of regions. Figure 5 visualizes the model layer-to-region mapping across the PFC surface. On its own, this mapping does not establish hierarchical processing, however, when combined with the hierarchy decoding accuracies and h_l values, we are able to project the model's hierarchical structure onto PFC ROIs (see Figures 6 & 7).

The decoding accuracy results in Figure 7 show that all model layers encode object-feature information, while only select layers additionally encode higher-level task-logic information. Regions including BA 8, 9, 46, and 47 exhibit a rostral–caudal progression of increased higher-level decoding accuracy aligning with the established theories of PFC hierarchy, though the progression is not smooth. Moreover, the most rostral regions (BA 10 and 11) and the most caudal regions (BA 6, IFJ, and IFS) deviate from the theorized rostral-caudal axis of hierarchy.

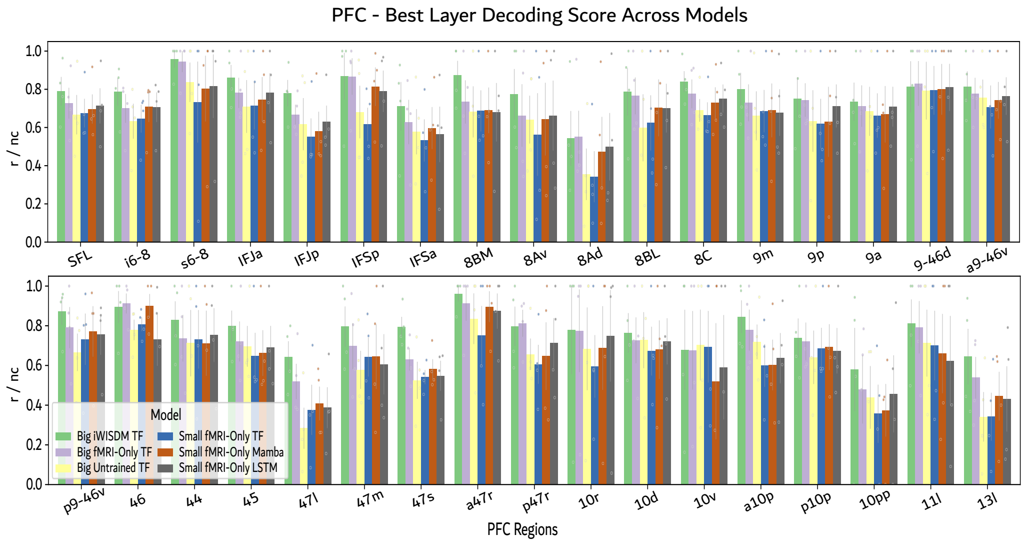


Figure 3: The best neural predictivity scores (NC corrected r-correlation) across model layers and regions of the PFC for every model tested.

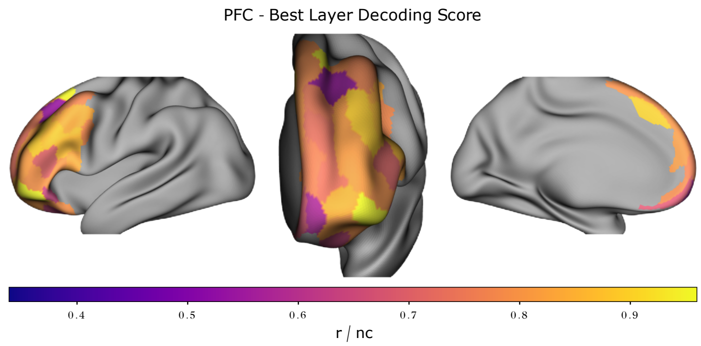


Figure 4: A PFC surface map of the best neural predictivity scores (NC corrected r-correlation) across layers of the Big iWISDM TF model.

The h_l hierarchy values in Figure 6 show that, rather than a clear rostral–caudal axis, results reveal a bullseye-like pattern centered on regions p47r, a9-46v, 9-46d, and 9a, with decreasing hierarchy values radiating outward in a ring-shape.

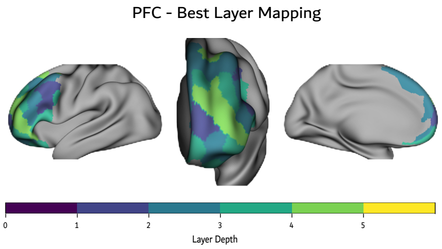


Figure 5: Big iWISDM TF model layer to PFC ROI surface map. Layers are selected based on their neural predictivity scores for each PFC ROI.

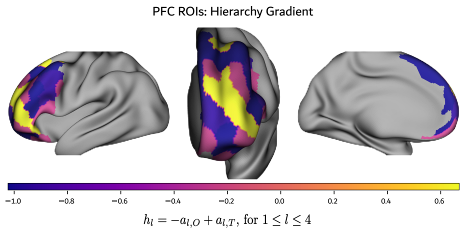

PFC ROIs: Hierarchy Gradient
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
$h_l = -a_{l,O} + a_{l,T}$, for $1 \leq l \leq 4$

Figure 6: Big iWISDM TF model layer hierarchy level (h_l) PFC surface plot. Layers are selected based on their neural predictivity scores for each PFC ROI.

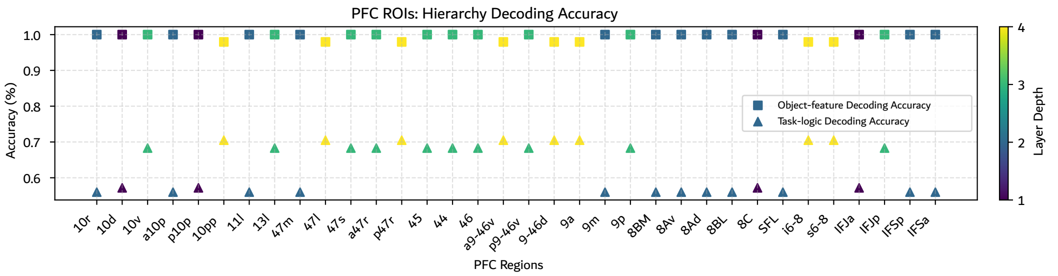


Figure 7: Big iWISDM TF model layer object-feature and task-logic decoding accuracy across PFC ROIs. Layers are selected based on their neural predictivity scores for each PFC ROI.

5 Discussion

In this work, we introduce a large multimodal transformer trained on millions of naturalistic decision-making and working memory tasks, and test its alignment with whole-brain fMRI recordings from humans performing the same tasks. Comparing model representations to neural activity at both network and regional prefrontal cortex levels, we find that model size and task generality strongly influence neural similarity. Using the resulting layer-to-region mappings, we further evaluate consistency with theories of hierarchical organization in the PFC.

5.1 Model size and generalization as key factors neural similarity

Our large, general-purpose Transformer outperforms baselines across cortical networks and PFC ROIs, consistent with prior work showing that ecologically trained models develop more brain-like representations [11, 24, 39, 41]. Both scale and training breadth matter, with our largest model, generalized across millions of working-memory tasks, showing the strongest alignment across higher-level networks and most PFC ROIs.

Input modality may also be contributing factor our observed model neural similarity results. The multi-task models were trained without language instructions, using one-hot task vectors, which may support the view that naturalistic inputs are needed for brain-like task representations; further experiments are required to disentangle this from task generalization.

Another surprising and notable finding was that the small Mamba2 model, despite having orders of magnitude fewer parameters, outperforms other small baselines and sometimes narrows the gap to larger models. Scaling and generalizing Mamba2 models is a promising direction for future work.

5.2 Insights into the hierarchical organization of the PFC

Beyond establishing neural similarity, our results provide novel insights into theories of hierarchical abstraction in the PFC. Using task feature decoding accuracies, we show that while all layers encode object-feature information, only select layers additionally capture higher-level task-logic information. Using these results, and mapping them onto the PFC, we find partial support for a rostral-caudal abstraction gradient, however, this gradient is not smooth and is not consistently expressed in the most rostral or caudal regions. Instead, a surface mapping of our hierarchy metric reveals a bullseye-like pattern, with a center in the centrolateral PFC ROIs.

5.3 Limitations and future directions

Several limitations of the present work should be noted:

- The fMRI dataset is currently unavailable to the public and will be made available upon publication.
- There is substantial variation in noise-ceiling estimates across subjects, which lowers the reliability of averaged predictivity scores.

- We did not compare against text-only baselines such as Llama or Centaur. Such comparisons would clarify the added value of high-resolution visual inputs relative to categorical textual inputs for predicting neural activity across the cortex.
- We did not evaluate general vision–language models fine-tuned on these tasks. Through training, these models may acquire more general and potentially more brain-like representations; testing them could refine conclusions about model class and training objectives.

Addressing these points is a priority for future work.

References

[1] Jie Lisa 2, Marjolein Spronk, Kaustubh Kulkarni, Grega Repovš, Alan Anticevic, and Michael W. Cole. Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*, 185:35–57, January 2019. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2018.10.006. URL `http://dx.doi.org/10.1016/j.neuroimage.2018.10.006`.

[2] Anonymous. Anonymous. *unpublished*, 185, December 2025. ISSN 1053-8119.

[3] Adam R. Aron, Trevor W. Robbins, and Russell A. Poldrack. Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, 8(4):170–177, April 2004. ISSN 1364-6613. doi: 10.1016/j.tics.2004.02.010. URL `http://dx.doi.org/10.1016/j.tics.2004.02.010`.

[4] Alan Baddeley. Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, 4(10):829–839, 2003. ISSN 14710048. doi: 10.1038/nrn1201. Number: 10.

[5] David Badre. Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, 12(5):193–200, May 2008. ISSN 1364-6613. doi: 10.1016/j.tics.2008.02.004. URL `https://www.sciencedirect.com/science/article/pii/S1364661308000612`.

[6] David Badre and Mark D'Esposito. Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of Cognitive Neuroscience*, 19(12): 2082–2099, December 2007. ISSN 0898-929X. doi: 10.1162/jocn.2007.19.12.2082.

[7] David Badre, Russell A. Poldrack, E. Juliana Paré-Blagoev, Rachel Z. Insler, and Anthony D. Wagner. Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. *Neuron*, 47(6):907–918, September 2005. ISSN 0896-6273. doi: 10.1016/j.neuron.2005.07.023. URL `http://dx.doi.org/10.1016/j.neuron.2005.07.023`.

[8] Jörg Bahlmann, Robert S. Blumenfeld, and Mark D'Esposito. The Rostro-Caudal Axis of Frontal Cortex Is Sensitive to the Domain of Stimulus Information. *Cerebral Cortex (New York, NY)*, 25(7):1815, January 2014. doi: 10.1093/cercor/bht419. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC4459285/`.

[9] Pouya Bashivan, Kohitij Kar, and James J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019. doi: 10.1126/science.aav9436.

[10] Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, July 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09215-4. URL `http://dx.doi.org/10.1038/s41586-025-09215-4`.

[11] Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, August 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09215-4. URL `https://www.nature.com/articles/s41586-025-09215-4`. Publisher: Nature Publishing Group.

[12] M BRASS, J DERRFUSS, B FORSTMANN, and D CRAMON. The role of the inferior frontal junction area in cognitive control. *Trends in Cognitive Sciences*, 9(7):314–316, July 2005. ISSN 1364-6613. doi: 10.1016/j.tics.2005.05.001. URL `http://dx.doi.org/10.1016/j.tics.2005.05.001`.

[13] Marcelo Camperi and Xiao-Jing Wang. A Model of Visuospatial Working Memory in Prefrontal Cortex: Recurrent Network and Cellular Bistability. *Journal of Computational Neuroscience*, 5(4):383–405, October 1998. ISSN 1573-6873. doi: 10.1023/A:1008837311948. URL `https://doi.org/10.1023/A:1008837311948`.

[14] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[15] Albert Compte, Nicolas Brunel, Patricia S. Goldman-Rakic, and Xiao Jing Wang. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex*, 10(9):910–923, 2000. ISSN 10473211. doi: 10.1093/cercor/10.9.910. Number: 9.

[16] Nelson Cowan. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, 104(2):163, 1988. ISSN 1939-1455. doi: 10.1037/0033-2909.104.2.163. URL `https://psycnet.apa.org/fulltext/1989-00336-001.pdf`. Publisher: US: American Psychological Association.

[17] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024. URL `https://arxiv.org/abs/2405.21060`.

[18] Laura N. Driscoll, Krishna Shenoy, and David Sussillo. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *Nature Neuroscience*, 27(7):1349–1363, July 2024. ISSN 1546-1726. doi: 10.1038/s41593-024-01668-6. URL `https://www.nature.com/articles/s41593-024-01668-6`. Publisher: Nature Publishing Group.

[19] P. S. Goldman-Rakic and P. Rakic. Preface: Cerebral cortex has come of age. *Cerebral Cortex*, 1(1):1–1, January 1991. ISSN 1460-2199. doi: 10.1093/cercor/1.1.1-a. URL `http://dx.doi.org/10.1093/cercor/1.1.1-a`.

[20] Ariel Goldstein et al. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022. doi: 10.1038/s41593-022-01026-4.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 1530-888X. doi: 10.1162/neco.1997.9.8.1735. URL `http://dx.doi.org/10.1162/neco.1997.9.8.1735`.

[22] Hyeon-Ae Jeon and Angela D. Friederici. Two principles of organization in the prefrontal cortex are cognitive hierarchy and degree of automaticity. *Nature Communications*, 4(1):2041, June 2013. ISSN 2041-1723. doi: 10.1038/ncomms3041. URL `https://www.nature.com/articles/ncomms3041`. Publisher: Nature Publishing Group.

[23] Jon H Kaas, Troy A Hackett, and Mark Jude Tramo. Auditory processing in primate cerebral cortex. *Current Opinion in Neurobiology*, 9(2):164–170, April 1999. ISSN 0959-4388. doi: 10.1016/s0959-4388(99)80022-1. URL `http://dx.doi.org/10.1016/s0959-4388(99)80022-1`.

[24] A. J. E. Kell, D. L. K. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644.e16, 2018. doi: 10.1016/j.neuron.2018.03.044.

[25] Etienne Koechlin, Gregory Corrado, Pietro Pietrini, and Jordan Grafman. Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning. *Proceedings of the National Academy of Sciences*, 97(13):7651–7656, June 2000. ISSN 1091-6490. doi: 10.1073/pnas.130177397. URL `http://dx.doi.org/10.1073/pnas.130177397`.

[26] Xiaoxuan Lei, Lucas Gomez, Hao Yuan Bai, and Pouya Bashivan. IWISDM: Assessing instruction following in multimodal models at scale, July 2024. URL http://arxiv.org/abs/2406.14343. arXiv:2406.14343 [cs].

[27] David E. Meyer and David E. Kieras. A computational theory of executive cognitive processes and multiple-task performance: Part 2. Accounts of psychological refractory-period phenomena. *Psychological Review*, 104:749–791, 1997. ISSN 1939-1471. doi: 10.1037/0033-295X.104.4.749. Place: US Publisher: American Psychological Association.

[28] Sabrina Namazova, Alessandra Brondetta, Younes Strittmatter, Matthew Nassar, and Sebastian Musslick. Not Yet AlphaFold for the Mind: Evaluating Centaur as a Synthetic Participant, August 2025. URL http://arxiv.org/abs/2508.07887. arXiv:2508.07887 [cs].

[29] Derek Evan Nee and Mark D'Esposito. The hierarchical organization of the lateral prefrontal cortex. *eLife*, 5:e12112, March 2016. ISSN 2050-084X. doi: 10.7554/eLife.12112. URL https://doi.org/10.7554/eLife.12112. Publisher: eLife Sciences Publications, Ltd.

[30] Derek Evan Nee and Mark D'Esposito. Causal evidence for lateral prefrontal cortex dynamics supporting cognitive control. *eLife*, 6:e28040, September 2017. ISSN 2050-084X. doi: 10.7554/eLife.28040. URL https://doi.org/10.7554/eLife.28040. Publisher: eLife Sciences Publications, Ltd.

[31] Randall C. O'Reilly, Todd S. Braver, and Jonathan D. Cohen. A Biologically Based Computational Model of Working Memory. In Akira Miyake and Priti Shah, editors, *Models of Working Memory*, pages 375–411. Cambridge University Press, 1 edition, April 1999. ISBN 978-0-521-58721-1 978-0-521-58325-1 978-1-139-17490-9. doi: 10.1017/CBO9781139174909.014. URL https://www.cambridge.org/core/product/identifier/CBO9781139174909A021/type/book_part.

[32] Michael Petrides. Lateral prefrontal cortex: architectonic and functional organization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):781–795, April 2005. ISSN 1471-2970. doi: 10.1098/rstb.2005.1631. URL http://dx.doi.org/10.1098/rstb.2005.1631.

[33] Emilia P. Piwek, Mark G. Stokes, and Christopher Summerfield. A recurrent neural network model of prefrontal brain activity during a working memory task. *PLOS Computational Biology*, 19(10):e1011555, 2023. doi: 10.1371/journal.pcbi.1011555.

[34] B. A. Richards, T. P. Lillicrap, P. Beaudoin, et al. A deep learning framework for neuroscience. *Nature Neuroscience*, 22:1761–1770, 2019. doi: 10.1038/s41593-019-0520-2.

[35] Frank E. Ritter, Farnaz Tehranchi, and Jacob D. Oury. ACT-R: A cognitive architecture for modeling cognition. *WIREs Cognitive Science*, 10(3):e1488, 2019. ISSN 1939-5086. doi: 10.1002/wcs.1488. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1488. _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.1488.

[36] 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? preprint, Neuroscience, September 2018. URL http://biorxiv.org/lookup/doi/10.1101/407007.

[37] Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. jina-embeddings-v3: Multilingual embeddings with task lora, 2024. URL https://arxiv.org/abs/2409.10173.

[38] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021. URL https://arxiv.org/abs/2104.00298.

[39] G. Tuckute, A. Sathe, S. Srikant, et al. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8:544–561, 2024. doi: 10.1038/s41562-023-01783-7.

[40] Zihan Weng, Lucas Gomez, Taylor Whittington Webb, and Pouya Bashivan. Caption This, Reason That: VLMs Caught in the Middle, May 2025. URL http://arxiv.org/abs/2505.21538. arXiv:2505.21538 [cs].

[41] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

[42] Guangyu Robert Yang, Igor Ganichev, Xiao-Jing Wang, Jonathon Shlens, and David Sussillo. A Dataset and Architecture for Visual Reasoning with a Working Memory, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Guangyu_Robert_Yang_A_dataset_and_ECCV_2018_paper.html.

[43] Guangyu Robert Yang, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):297–306, 2019. doi: 10.1038/s41593-018-0310-2.

[44] Guangyu Robert Yang, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):297–306, February 2019. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-018-0310-2. URL http://www.nature.com/articles/s41593-018-0310-2. Number: 2.

A Appendix

A.1 Surface maps

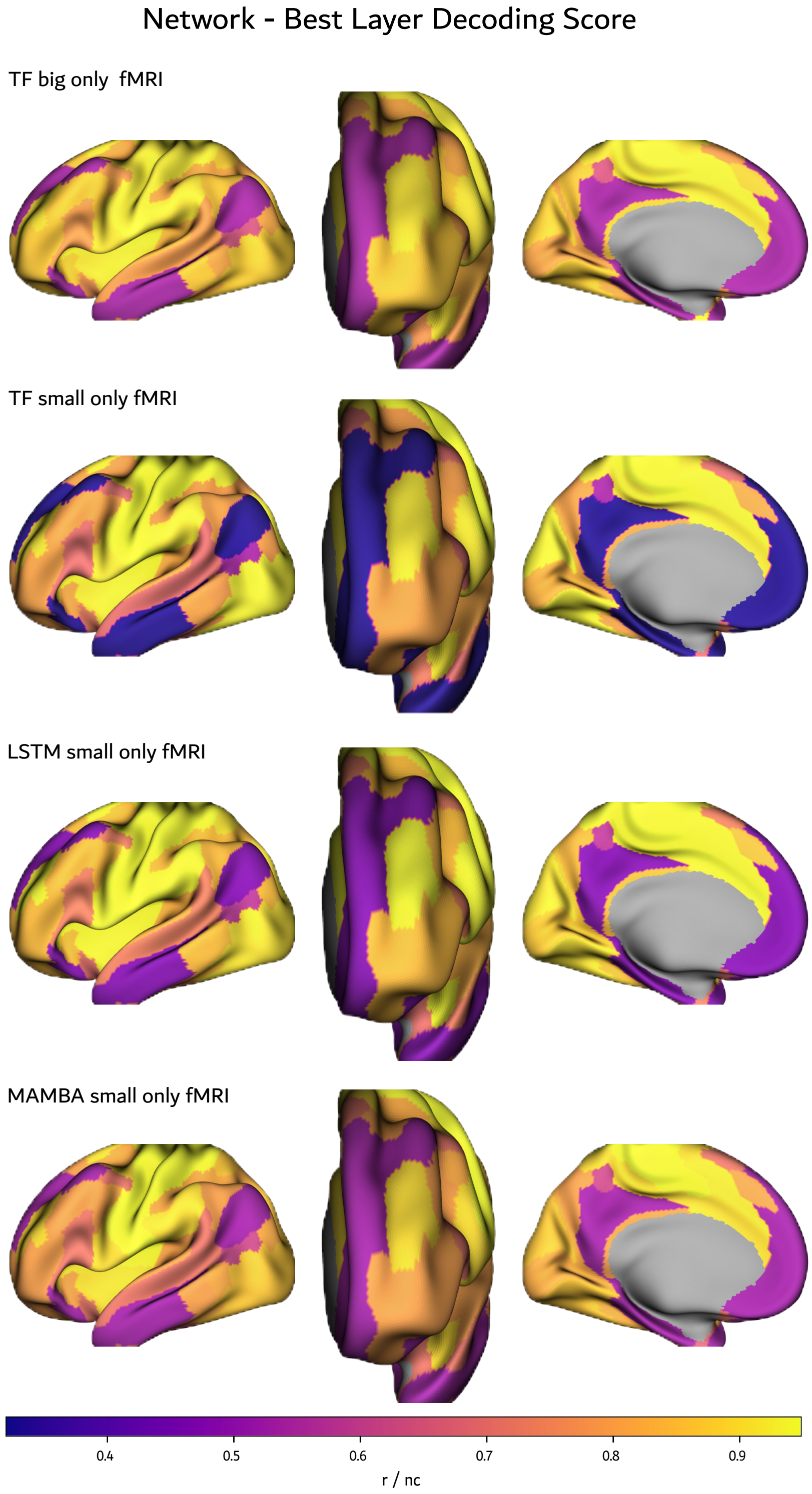


Figure 8: A network surface map of the best neural predictivity scores (NC corrected r-correlation) across layers for all models

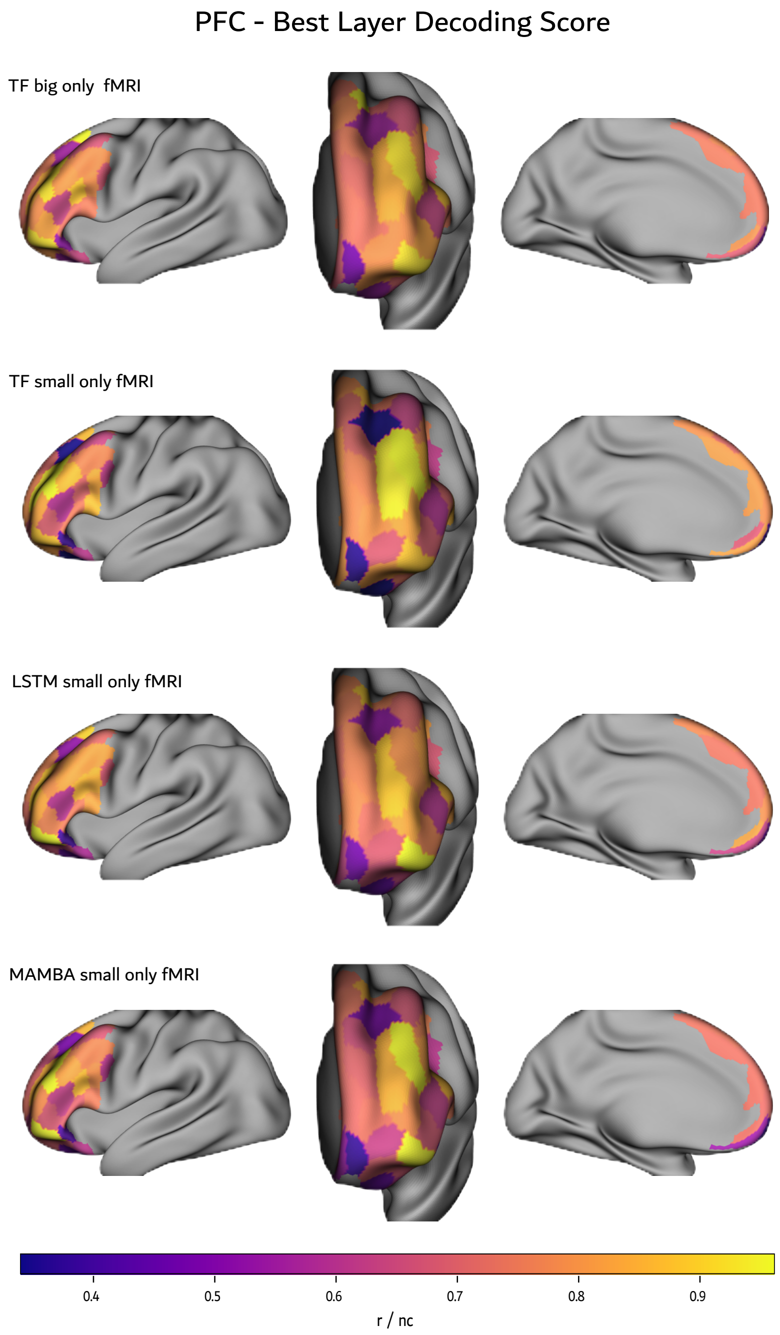
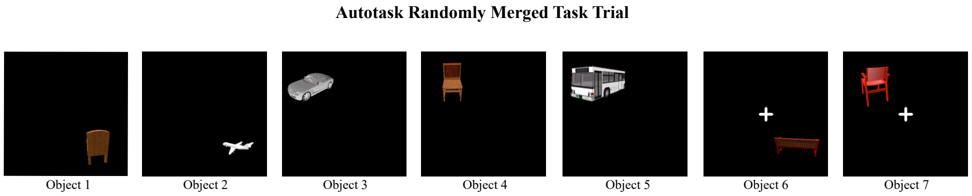


Figure 9: A PFC surface map of the best neural predictivity scores (NC corrected r-correlation) across layers for all models.

A.2 iWISDM and fMRI data examples



Observe object 1, observe object 2, observe object 3, observe object 4, observe object 5, observe object 6,
if location of object 3 equals location of object 5 and identity of object 4 not equals identity of object 5, then category of object 2 not equals category of object 1?
else benches not equals category of object 6? observe object 7, if location of object 5 not equals location of object 6, then location of object 7?
else category of object 3 not equals category of object 4?
Answer: Null, Null, Null, Null, Null, True, **Top-Left**

Figure 10: Example of an iWISDM rendered task trial

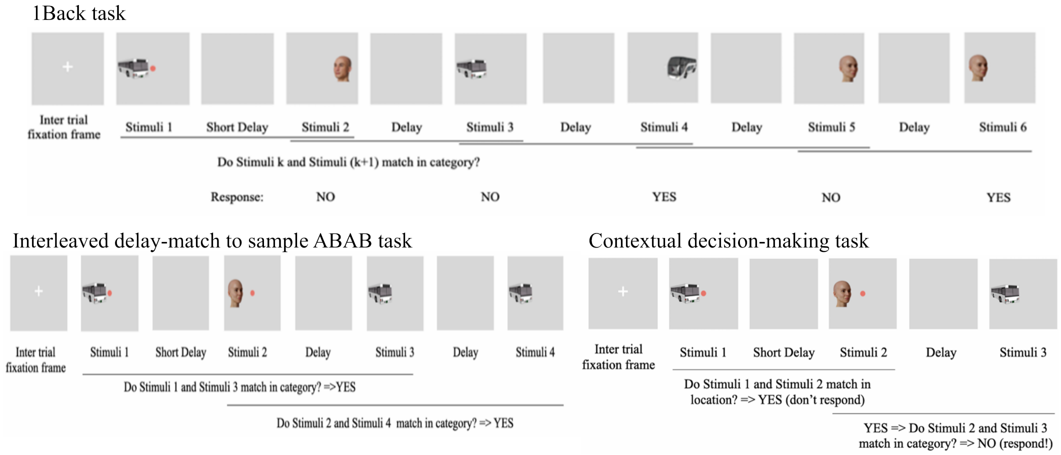

1Back task
Inter trial fixation frame
Stimuli 1
Short Delay
Stimuli 2
Delay
Stimuli 3
Delay
Stimuli 4
Delay
Stimuli 5
Delay
Stimuli 6
Do Stimuli k and Stimuli (k+1) match in category?
Response:
NO
NO
YES
NO
YES
Interleaved delay-match to sample ABAB task
Inter trial fixation frame
Stimuli 1
Short Delay
Stimuli 2
Delay
Stimuli 3
Delay
Stimuli 4
Do Stimuli 1 and Stimuli 3 match in category? =>YES
Do Stimuli 2 and Stimuli 4 match in category? => YES
Contextual decision-making task
Inter trial fixation frame
Stimuli 1
Short Delay
Stimuli 2
Delay
Stimuli 3
Do Stimuli 1 and Stimuli 2 match in location? => YES (don't respond)
YES => Do Stimuli 2 and Stimuli 3 match in category? => NO (respond!)

Figure 11: Example of an fMRI rendered task trial

A.3 Model architecture and hyperparameter

Table 1: Model configuration and training hyperparameters.

Component	Setting	Training	Setting
Image encoder	EfficientNet	Effective batch size	1400
Instruction encoder	Jina Embeddings v3	Trials per epoch	14,000
Core decoder	Transformer	Epochs	1500
FFL dimension	9216	Warmup steps	600
Activation	SwiGLU	Peak learning rate	3.5e-5
Hidden dimension	1536	Min learning rate	1.0e-7
Number of heads	16	Scheduler	HF Cosine
Number of blocks	6	GPUs	4 × H100
Dropout	0.1		
Weight decay	0.1		