arXiv:2507.01966v1 [q-bio.NC] 18 Jun 2025

Alignment between Brains and AI: Evidence for Convergent Evolution across Modalities, Scales and Training Trajectories

Guobin Shen[1,3,†], **Dongcheng Zhao**[1,4,†], **Yiting Dong**[1,3], **Qian Zhang**[1,4], **and Yi Zeng**[1,2,4,*]

[1]Laboratory of Brain-inspired Cognitive AI, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
[2]State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Chinese Academy of Sciences, Shanghai, 200031, China
[3]School of Future Technology, University of Chinese Academy of Sciences, Beijing, 100049, China
[4]Center for Long-term AI, Beijing, 101407, China
[*]corresponding author: Yi Zeng (yi.zeng@ia.ac.cn)
[†]these authors contributed equally to this work

ABSTRACT

Artificial and biological systems may evolve similar computational solutions despite fundamental differences in architecture and learning mechanisms—a form of convergent evolution. We demonstrate this phenomenon through large-scale analysis of alignment between human brain activity and internal representations of over 600 AI models spanning language and vision domains, from 1.33M to 72B parameters. Analyzing 60 million alignment measurements reveals that higher-performing models spontaneously develop stronger brain alignment without explicit neural constraints, with language models showing markedly stronger correlation ($r = 0.89$, $p < 7.5 \times 10^{-13}$) than vision models ($r = 0.53$, $p < 2.0 \times 10^{-44}$). Crucially, longitudinal analysis demonstrates that brain alignment consistently precedes performance improvements during training, suggesting that developing brain-like representations may be a necessary stepping stone toward higher capabilities. We find systematic patterns: language models exhibit strongest alignment with limbic and integrative regions, while vision models show progressive alignment with visual cortices; deeper processing layers converge across modalities; and as representational scale increases, alignment systematically shifts from primary sensory to higher-order associative regions. These findings provide compelling evidence that optimization for task performance naturally drives AI systems toward brain-like computational strategies, offering both fundamental insights into principles of intelligent information processing and practical guidance for developing more capable AI systems.

Introduction

The rapid progress of Artificial Intelligence (AI) has raised a fundamental question: as these systems increasingly match or surpass human-level performance in language, vision, and reasoning domains[1–3], do their internal representations also converge toward brain-like computational strategies? This question becomes more pressing as AI-generated outputs become indistinguishable from human-created content in interactive settings[4]. Understanding these internal representational mechanisms is increasingly urgent both for advancing our knowledge of intelligence and for guiding the development of safer, more interpretable AI systems.

Recent advances in Large Language Models (LLMs)[3,5,6] and vision models[7,8] provide an unprecedented opportunity to empirically test this convergent evolution hypothesis. Prior studies have demonstrated that deep neural networks map onto primate visual cortex hierarchies[9] or transformer-based LLMs align with human language areas[10]. However, existing work has been constrained by three key limitations: analyses restricted to single modalities, examination of narrow cortical regions rather than distributed networks, and reliance on static model checkpoints that fail to capture representational evolution during training[11,12]. These constraints have prevented a comprehensive understanding of whether and how AI systems converge toward brain-like representations across diverse computational domains.

In both biological and artificial systems, internal representations refer to distributed patterns of activity that encode information relevant for behavior and cognition. In the brain, these representations are shaped by experience and task demands, forming hierarchical organizations from sensory to conceptual levels[13]. Deep learning models exhibit analogous properties, with internal states evolving across layers to capture increasingly abstract features. Examining these internal dynamics beyond merely analyzing outputs uncovers fundamental principles of intelligent information processing. This approach illuminates the

mechanisms underlying the emergence of brain-like computational strategies, providing insights into the fundamental nature of intelligence.

Convergent evolution provides a powerful framework for understanding brain-AI alignment. Just as vertebrates and cephalopods independently evolved camera-like eyes under similar visual processing demands[14,15], artificial and biological systems may converge on similar computational strategies when facing shared information-processing challenges[16–18]. This convergence has profound implications: high-performing AI models offer testbeds for understanding biological computation[18,19], while brain organization may guide development of more robust and efficient AI systems[16]. Brain-AI alignment thus serves as both a diagnostic tool and a blueprint for future intelligence architectures.

To systematically test this convergent evolution hypothesis and address the limitations of prior work, we present a comprehensive analysis of brain-AI alignment unprecedented in scale. Our study examines internal representations from over 600 models across diverse architectures, scales, training trajectories, and modalities. Through analyzing 60 million alignment measurements, we directly compare layer-wise activations to human neural recordings, addressing three fundamental questions: (1) Does brain alignment precede performance improvements during training, suggesting it may be a necessary stepping stone? (2) Do alignment patterns differ systematically between vision and language modalities, or converge toward universal principles? (3) How do model hierarchies progressively map onto cortical processing levels from sensory to associative regions? Our findings reveal that artificial and biological intelligence, despite their distinct evolutionary paths, indeed converge toward similar computational solutions. This establishes fundamental principles that could transform both our understanding of intelligence and the development of future AI systems.

Results

We developed a systematic large-scale framework to comprehensively assess brain-AI representational alignment across sensory modalities, model scales, and training dynamics (Figure 1). Our analysis leveraged fMRI recordings from the Natural Scenes Dataset (NSD)[20], which captures neural activity from multiple subjects viewing thousands of naturalistic images. These images, sourced from the COCO dataset[21] and paired with human-generated captions, enabled multimodal analyses across both vision and language domains.

Our AI model collection comprised 630 neural networks spanning diverse architectures and scales: 36 large language models (0.5-72B parameters) including Qwen[6], Llama[5], and Gemma[22] families, and 594 vision models (1.33-1014M parameters) from CNNs to transformers[7,23]. To quantify alignment, we employed Centered Kernel Alignment (CKA)[24] across multiple spatial scales, mapping model representations to brain regions defined by the HCP_MMP1 parcellation[25] and Yeo-7 functional networks[26]. For longitudinal analyses, we tracked representational evolution in the Pythia language model family[27] and MixNet vision model family[28] throughout training.

Correlation Patterns Between Brain Alignment and Model Effectiveness

Our first key finding is a robust, positive correlation between model performance and brain alignment across both language and vision domains (Figure 1d-e). For performance evaluation, we utilized two widely recognized benchmarks: the LLM Leaderboard 2[29] for language models, which aggregates performance across multiple reasoning, knowledge, and language understanding tasks into a composite score, and ImageNet[30] Top-1 accuracy for vision models, which measures classification performance on a diverse dataset of 1,000 object categories. For language models, brain alignment showed a strong correlation with performance (Pearson $r = 0.89$, $p < 7.5 \times 10^{-13}$; Spearman $\rho = 0.88$, $p < 2.7 \times 10^{-12}$), with the relationship better characterized by a logarithmic fit ($R^2 = 0.80$, $AIC = 117.6$, $BIC = 120.8$) than a linear one ($R^2 = 0.78$, $AIC = 120.7$, $BIC = 123.9$). This pattern held across different model families, with higher-performing models generally showing stronger brain alignment.

The logarithmic relationship suggests a principled trajectory of diminishing returns, where initial improvements in performance correspond to substantial gains in brain alignment, followed by more modest alignment improvements as performance continues to increase. Notably, we observed that language models with similar performance but different architectural designs and fine-tuning approaches showed distinct brain alignment scores. For instance, Llama-3.1-Tulu-3-8B[31] exhibits higher brain alignment and performance than the base Llama-3.1-8B[5] model, suggesting that certain fine-tuning approaches may enhance brain-like representations. Similarly, gemma-2-9b-SimPO[32] shows lower brain alignment and performance compared to the standard gemma-2-9b[22], indicating that different training methodologies impact alignment with biological representations even within the same model family.

Similarly, for vision models evaluated on ImageNet, we observed a significant correlation between accuracy and brain alignment (Pearson $r = 0.53$, $p < 2.0 \times 10^{-44}$; Spearman $\rho = 0.47$, $p < 2.5 \times 10^{-33}$). The correlation was consistent across model sizes, from small models (under 10M parameters) to large models (over 1000M parameters), indicating a scale-invariant relationship between performance optimization and brain alignment. An important observation in the vision domain is that

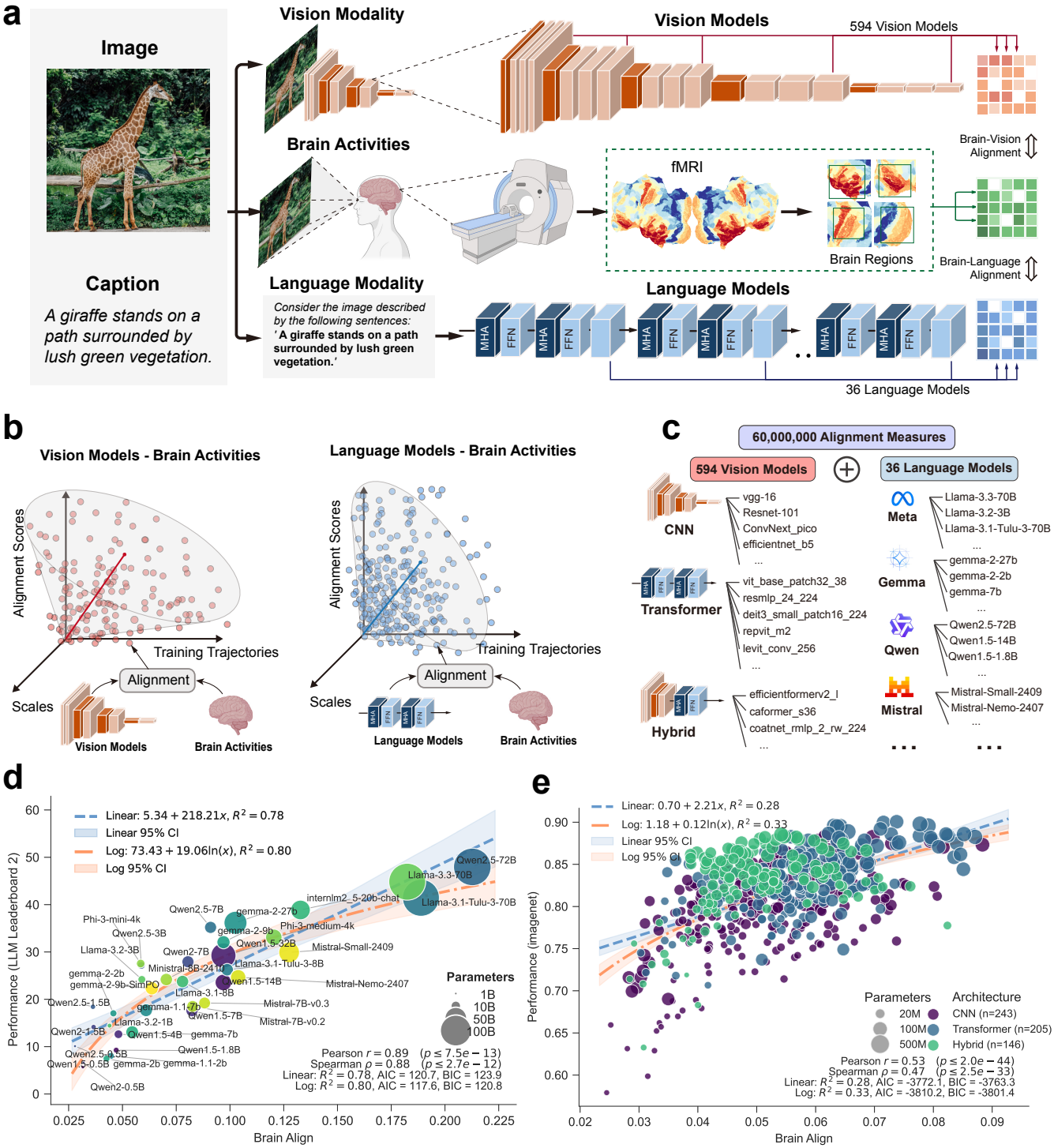

Figure 1. Relationship between model performance and brain alignment. a, Methodological overview showing the multi-modal analysis framework for vision models, brain recordings, and language models. **b**, Multi-dimensional analytical approach spanning scales, training trajectories, and modalities for Brain-AI alignment. **c**, Comprehensive analysis of 594 vision models and 36 language models, enabling 60 million distinct alignment measurements. **d**, Language models show strong correlation between LLM Leaderboard 2[29] performance and brain alignment scores, with logarithmic fit outperforming linear fit. Parameter counts indicated by marker size. **e**, Vision models demonstrate moderate but significant correlation between ImageNet top-1 accuracy and brain alignment, with differentiation between architectural types.

transformer-based vision models exhibited slightly higher brain alignment than CNNs at equivalent performance levels, suggesting that self-attention mechanisms may naturally encourage more brain-like representations.

Both vision and language models exhibit a logarithmic relationship between performance and brain alignment, revealing an

interesting pattern of diminishing returns. As model performance approaches benchmark saturation, brain alignment scores continue to increase. This suggests models may develop increasingly brain-like representations even after task performance reaches ceiling effects on standard benchmarks.

Model Information		Pearson Correlation				Spearman Correlation			
Model	**Benchmark**	r	**95% CI**	p**-value**	**FDR p**	ρ	**95% CI**	p**-value**	**FDR** p
Language Models	IFEval[33]	0.77	[0.59, 0.88]	<3.5e-08	<4.1e-08	0.79	[0.58, 1.00]	<1.3e-08	<1.6e-08
	BBH[34]	0.84	[0.70, 0.91]	<2.3e-10	<3.0e-10	0.86	[0.65, 1.07]	<1.9e-11	<2.7e-11
	MATH Lvl 5[35]	0.76	[0.58, 0.87]	<7.6e-08	<8.7e-08	0.67	[0.46, 0.88]	<7.1e-06	<7.6e-06
	GPQA[36]	0.76	[0.57, 0.87]	<7.7e-08	<8.7e-08	0.78	[0.57, 0.99]	<2.3e-08	<2.8e-08
	MUSR[37]	0.66	[0.42, 0.81]	<1.3e-05	<1.3e-05	0.59	[0.38, 0.80]	<1.6e-04	<1.6e-04
	MMLU-PRO[38]	0.83	[0.69, 0.91]	<3.8e-10	<5.0e-10	0.85	[0.64, 1.06]	<7.3e-11	<1.0e-10
	Leaderboard 2[29]	0.89	[0.79, 0.94]	<7.5e-13	<1.1e-12	0.88	[0.67, 1.09]	<2.7e-12	<4.0e-12
	Chatbot Arena[39]	0.73	[0.48, 0.87]	<2.3e-05	<2.4e-05	0.80	[0.55, 1.05]	<9.9e-07	<1.1e-06
Vision Models	ImageNet[30]	0.53	[0.47, 0.59]	<2.0e-44	<1.5e-43	0.47	[0.41, 0.52]	<2.5e-33	<6.5e-33
	ImageNet-A[40]	0.46	[0.39, 0.52]	<7.5e-32	<1.7e-31	0.45	[0.40, 0.50]	<2.1e-31	<4.4e-31
	ImageNet-A-Clean[40]	0.51	[0.45, 0.57]	<1.7e-41	<7.0e-41	0.46	[0.41, 0.51]	<2.3e-32	<5.6e-32
	ImageNet-R[41]	0.52	[0.46, 0.58]	<8.3e-43	<4.6e-42	0.47	[0.42, 0.52]	<1.1e-33	<3.1e-33
	ImageNet-R-Clean[40]	0.51	[0.45, 0.57]	<2.8e-40	<1.0e-39	0.45	[0.40, 0.50]	<5.5e-31	<1.1e-30
	ImageNet-Real[42]	0.50	[0.44, 0.56]	<1.2e-38	<4.0e-38	0.44	[0.39, 0.49]	<2.9e-29	<5.0e-29
	ImageNetv2-matched-freq.[43]	0.52	[0.46, 0.57]	<5.6e-42	<2.4e-41	0.45	[0.40, 0.50]	<1.3e-30	<2.4e-30
	sketch[44]	0.53	[0.47, 0.58]	<1.3e-43	<8.2e-43	0.47	[0.42, 0.52]	<4.0e-34	<1.2e-33

Table 1. Correlation between model performance and brain alignment across benchmarks. Results show correlation metrics across different performance benchmarks for language and vision models. FDR-corrected p-values account for multiple comparisons.

Table 1 presents detailed correlation statistics between brain alignment and performance across multiple benchmarks. For language models, the composite LLM Leaderboard 2 metric showed the strongest correlation ($r = 0.89$), with component benchmarks exhibiting similarly robust relationships: BBH ($r = 0.84$), MMLU-PRO ($r = 0.83$), IFEval ($r = 0.77$), MATH Level 5 ($r = 0.76$), and GPQA ($r = 0.76$). Notably, Chatbot Arena[39], which relies on human evaluation of conversational quality, showed a slightly lower correlation ($r = 0.73$). This difference likely reflects that our measurement paradigm aligns more closely with objective capability assessment than with the subjective criteria underlying human conversational preferences. Vision models demonstrated moderate but consistent correlations across evaluation contexts, from standard ImageNet ($r = 0.53$) to more challenging variants like ImageNet-R ($r = 0.52$) and ImageNet-A ($r = 0.46$). This pattern suggests that higher-level cognitive tasks, particularly those requiring complex reasoning abilities, may more strongly drive representational convergence between artificial and biological systems. Additional visualizations showing performance-brain alignment correlations for all metrics listed in Table 1 are provided in the supplementary materials.

The consistent brain-performance correlation across vastly different model architectures, training objectives, and modalities provides compelling evidence for a form of convergent evolution, whereby optimization for task performance naturally leads AI systems toward more brain-like representations without explicit neurobiological constraints. The substantially stronger correlations observed in language models compared to vision models likely reflects that language processing involves more abstract and complex cognitive operations, such as semantic integration, contextual reasoning, and compositional understanding, which may naturally drive both biological and artificial systems toward similar computational solutions. This modality gap in correlation strength is maintained across different benchmarks and evaluation contexts, suggesting that as tasks become more cognitively demanding, artificial systems increasingly converge with biological processing strategies.

Importantly, these alignment patterns were highly consistent across different subjects, as demonstrated by the strong inter-subject correlations in both vision and language modalities. Our comprehensive consistency analysis verified this robustness (Tables 2, 3), with inter-subject correlations ranging from 0.997 to 1.000 for language models and 0.970 to 0.999 for vision

models. These remarkably high cross-subject correlations indicate that the observed representational correspondences reflect universal computational principles rather than individual variations, suggesting that both biological and artificial intelligence systems converge on similar solutions when facing the same information processing challenges.

Architectural Correspondences Between Brains and AI

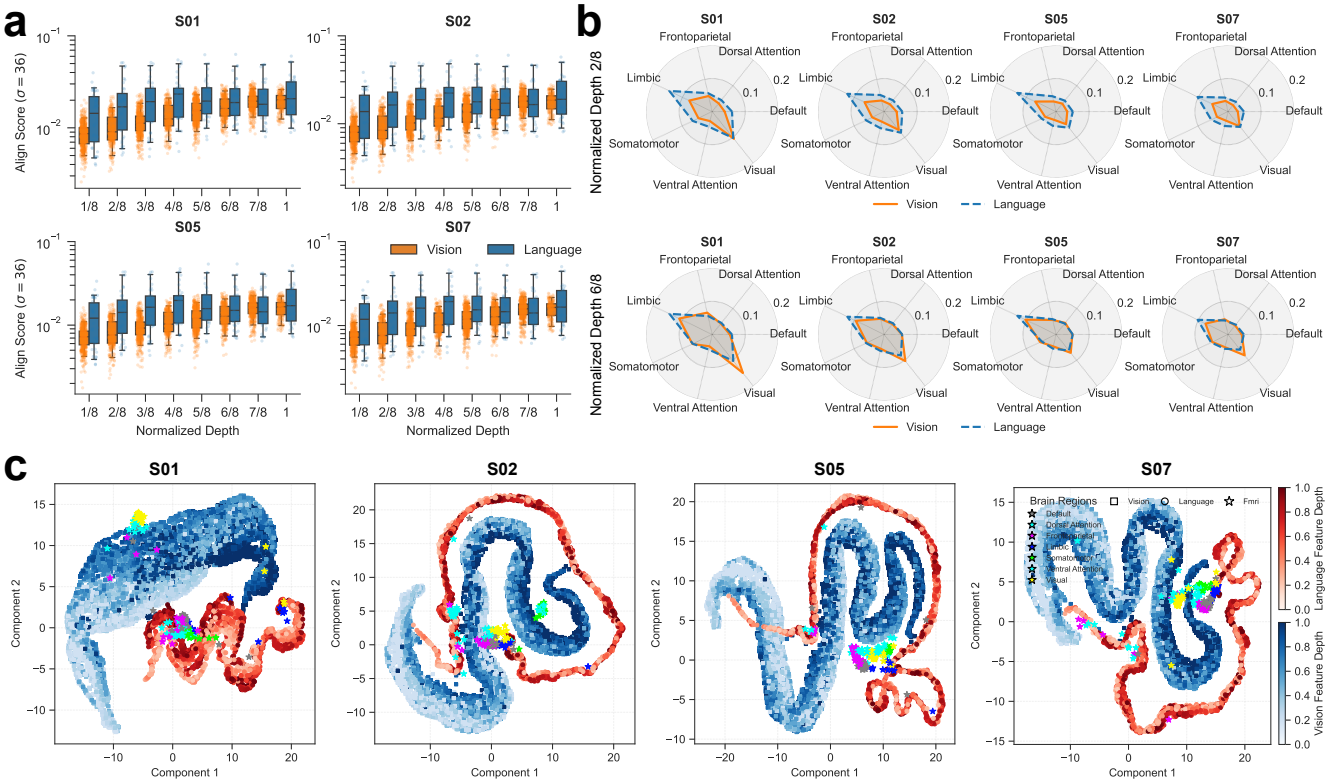

Figure 2. Layer-wise brain alignment patterns across vision and language models. a, Brain alignment by normalized layer depth for vision and language models across subjects. For consistent comparison across models with different architectures, each model's layers were normalized into 8 depth segments (1/8 through 8/8). **b**, Comparison of brain network alignment at normalized depths 2/8 and 6/8 for vision versus language models across Yeo-7 brain networks[26]. **c**, UMAP[45] visualization of 2D embedding space based on brain region alignment patterns, showing relationships between model layers and brain regions. Each point represents either a model layer or brain region, with color intensity indicating the depth for model layers, and different colors representing distinct brain regions.

We next examined how brain alignment varies across layers of deep neural networks. To enable consistent comparisons across architectures of varying depth, we divided each model into 8 normalized layer segments (from 1/8 to 8/8) and averaged alignment scores within each. As shown in Figure 2a, vision models exhibited a gradual increase in alignment with depth, peaking at the final segments (6/8 to 8/8). In contrast, language models showed overall higher brain alignment, with a distinct middle-layer peak (3/8 to 5/8) that remained relatively stable through deeper layers.

When examining alignment with specific brain networks defined by the Yeo-7[26] parcellation (Figure 2b), we observed clear modality-specific patterns. At normalized depth 2/8 (shallow layers), vision models showed modest alignment across all networks with a slight preference for visual regions, while language models demonstrated stronger overall alignment with pronounced correspondence to limbic and default mode networks. At normalized depth 6/8 (deeper layers), vision models developed substantially stronger alignment with visual network regions while maintaining lower alignment to other networks. Language models at this deeper level showed a more balanced pattern across networks, maintaining strong alignment to limbic and default regions. This distinct pattern was consistent across all subjects, suggesting a fundamental organizational principle rather than individual variation.

Interestingly, language models showed substantial alignment with non-language-specific brain regions associated with abstract concept representation and semantic integration. This pattern suggests that language models may capture general-purpose computational principles employed across diverse cognitive domains rather than language-specific representations

alone. Vision models, conversely, showed more modality-specific alignment concentrated in visual processing regions.

Dimensionality reduction analysis further illuminated these representational relationships (Figure 2c). Using UMAP[45] to project the high-dimensional alignment scores into a 2D space revealed two distinct manifolds: vision (blue squares) and language (red circles) model representations, with color intensity indicating layer depth progression from early to deeper processing stages. Brain regions (represented as stars) are derived from the HCP_MMP1[25] parcellation and are colored according to their Yeo-7 network affiliation. These brain regions are distributed across the space based on their alignment patterns with different model layers, providing a visual map of how neural and artificial representations relate to each other.

Notably, deeper vision model layers positioned closer to the language model manifold, suggesting that as visual processing becomes more abstract, it begins to share representational properties with language processing. Brain regions followed a similar organizational pattern, with visual cortex regions (yellow stars) clustering near early and middle vision model layers, while association cortices and limbic regions (blue and purple stars) positioned closer to language model representations. This visualization provides striking evidence for a hierarchical organization of representational alignment spanning both biological and artificial systems, with deeper processing stages across modalities converging toward similar representational strategies for high-level information integration.

Brain Alignment as a Precursor to Performance

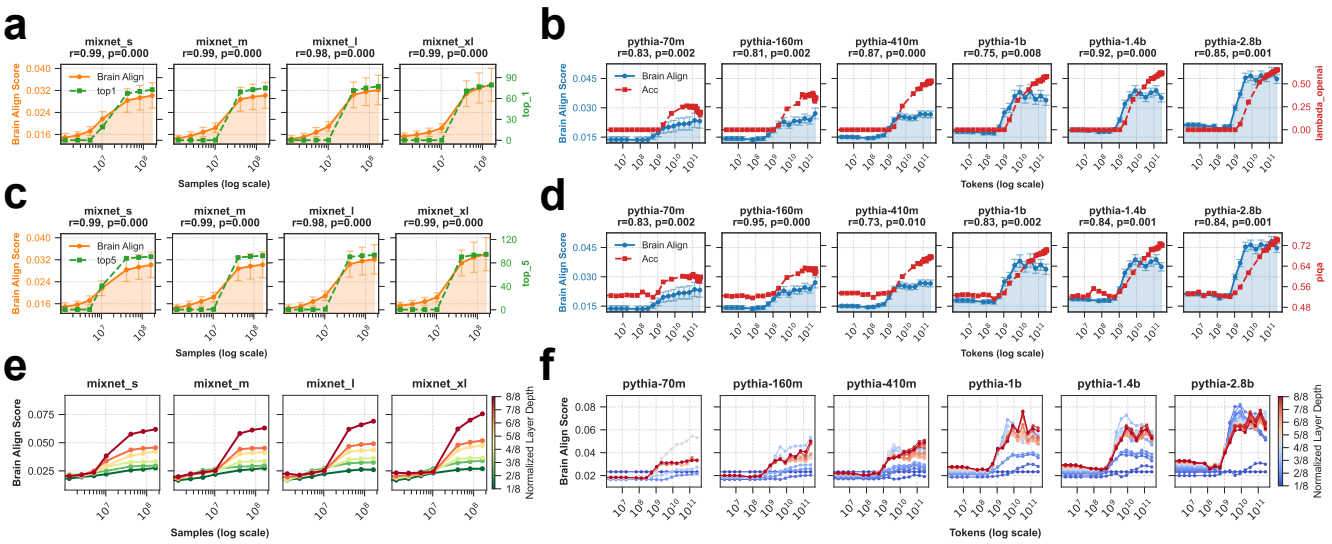

Figure 3. Evolution of brain alignment during model training. a-d, Dual y-axis plots showing brain alignment score (left axis) and performance metrics (right axis) throughout training for vision models (MixNet family[28]) and language models (Pythia family[27]). **e-f**, Line plots showing layer-wise brain alignment evolution across training progression for vision and language models.

While our previous analyses established a robust correlation between model performance and brain alignment across different modalities and architectures, the temporal relationship between these variables remains unexplored. Understanding whether brain alignment precedes performance improvements—or vice versa—is crucial for distinguishing between two competing hypotheses: performance-driven convergence, where optimization for task success naturally leads to brain-like representations; and alignment-driven performance, where developing brain-like representations serves as a necessary stepping stone toward higher capabilities. Clarifying this temporal relationship could reveal whether the convergent evolution between artificial and biological systems follows a predictable developmental trajectory.

Although longitudinal neuroimaging studies in humans are constrained because development unfolds over years and data collection is resource-intensive, artificial neural networks provide a unique experimental paradigm. Saved checkpoints across training offer high-resolution temporal snapshots of representational evolution, thus enabling direct testing of whether brain alignment emerges as a precursor, a consequence, or a concurrent feature of performance improvements.

Longitudinal analysis of model training revealed that brain alignment is not merely a static property but evolves dynamically throughout the learning process (Figure 3a-d). For both vision models (MixNet family)[28] and language models (Pythia family)[27], we tracked brain alignment alongside performance metrics (ImageNet[30] top-1 and top-5 accuracy for vision models; LAMBADA[46] and PIQA[47] scores for language models) across training checkpoints.

A critical finding emerged from this analysis: increases in brain alignment consistently preceded improvements in task performance. For vision models, brain alignment increased throughout the entire training process, with particularly sharp rises in early training (reaching approximately 85% of maximum alignment by just 20% of training samples), while performance only began improving after approximately 5% of training and continued to increase more gradually thereafter. The correlation between brain alignment and performance across the entire training process was strong ($r = 0.98$ for MixNet-L). This pattern suggests that developing brain-like representations may be a necessary precursor to achieving high performance rather than merely a consequence of performance improvement.

This leading indicator relationship was even more pronounced in language models, where brain alignment began increasing as early as 1% of training and rapidly reached stable levels by approximately 10% of training tokens. In stark contrast, performance metrics on downstream tasks such as PIQA and LAMBADA showed significant delays, only reaching half of their maximum values after approximately 10% of training. Across all Pythia models, we observed consistently strong correlations between brain alignment and task performance (ranging from $r = 0.73$ to $r = 0.95$, all $p < 0.01$), with brain alignment increases consistently preceding performance improvements. This pattern held across model scales, reinforcing the hypothesis that developing brain-like representations may serve as a fundamental precursor to performance gains in language models.

The stabilization of overall brain alignment around 10% of training is further elucidated by a layer-wise analysis, revealing distinct, modality-specific patterns (Figure 3e-f). In vision models, brain alignment steadily increased across all layers throughout training, with deeper layers ultimately achieving higher alignment scores than earlier layers.

In contrast, language models exhibited a different developmental trajectory. Early layers (normalized depth 1/8 to 2/8) maintained relatively stable and low alignment levels throughout training. This limited early-layer alignment can likely be attributed to the nature of the NSD dataset, which primarily features visual stimuli rather than rich linguistic content, offering little opportunity for alignment with token-level representations.

Middle layers (normalized depth 4/8 to 6/8) showed a rise-then-fall pattern: alignment initially increased during early training (up to 10% of training tokens), followed by a moderate decline as training progressed. The deeper layers (normalized depth 7/8 to 8/8) exhibited a delayed growth pattern, with alignment continuing to increase even after middle layers had begun to decline, eventually stabilizing with minor fluctuations.

These layer-specific dynamics became increasingly pronounced in larger models. Models such as pythia-1b, pythia-1.4b, and pythia-2.8b exhibited more dramatic differentiation between early, middle, and deep layer behaviors compared to their smaller counterparts. This scale-dependent pattern suggests that larger models not only develop stronger brain alignment overall but also evolve more specialized and functionally differentiated internal representations across layers.

These training dynamics reinforce a central aspect of our convergent evolution hypothesis: as models become more effective at solving their target tasks, they naturally discover representational strategies that align with biological solutions, even in the absence of explicit constraints promoting brain-like processing. The observation that alignment precedes performance suggests that brain-like representations may be stepping stones toward effective task solutions rather than mere by-products of optimization.

Multi-scale Brain Alignment Reveals Regional Processing Hierarchies

Previous research comparing brain activity patterns with AI representations has typically relied on single alignment metrics with fixed parameters, without systematically investigating how measurement scale impacts the observed correspondence between artificial and biological systems. While our analyses demonstrated significant consistency across different measures (Figure 6), the choice of measurement scale may nonetheless reveal important organizational principles in both systems that would otherwise remain hidden.

To explore this possibility, we systematically varied the kernel size in our CKA measurements from smaller to larger scales, uncovering a striking shift in which brain regions showed strongest model correspondence (Figure 4a). This methodological approach has direct biological significance: smaller kernel sizes (lower σ values) focus on very similar activation patterns across stimuli, capturing highly specific representational similarities analogous to how primary sensory areas respond selectively to specific features. In contrast, larger kernel sizes (higher σ values) detect similarities across more varied stimuli, reflecting broader representational patterns similar to how association cortices integrate diverse information across contexts.

Using the HCP_MMP1[25] parcellation grouped according to the Yeo-7 network taxonomy[26], we observed that as kernel size increased, primary visual areas showed substantial decreases in relative alignment, while limbic, ventral attention, and default mode networks showed increasing alignment.

The most dramatic decreases in alignment with increasing kernel size were observed in early visual areas V1 ($\Delta = -5.30$), V2 ($\Delta = -2.81$), and V3 ($\Delta = -1.05$), while the strongest increases occurred in limbic regions including orbital frontal cortex (OFC, $\Delta = +1.56$), temporal regions TF ($\Delta = +1.36$), TE2a ($\Delta = +1.36$), and TGv ($\Delta = +1.33$), and parahippocampal regions EC ($\Delta = +1.09$) and PeEc ($\Delta = +1.06$).

These shifts were remarkably consistent across subjects, with a sign consistency of 0.83 in normalized alignment changes

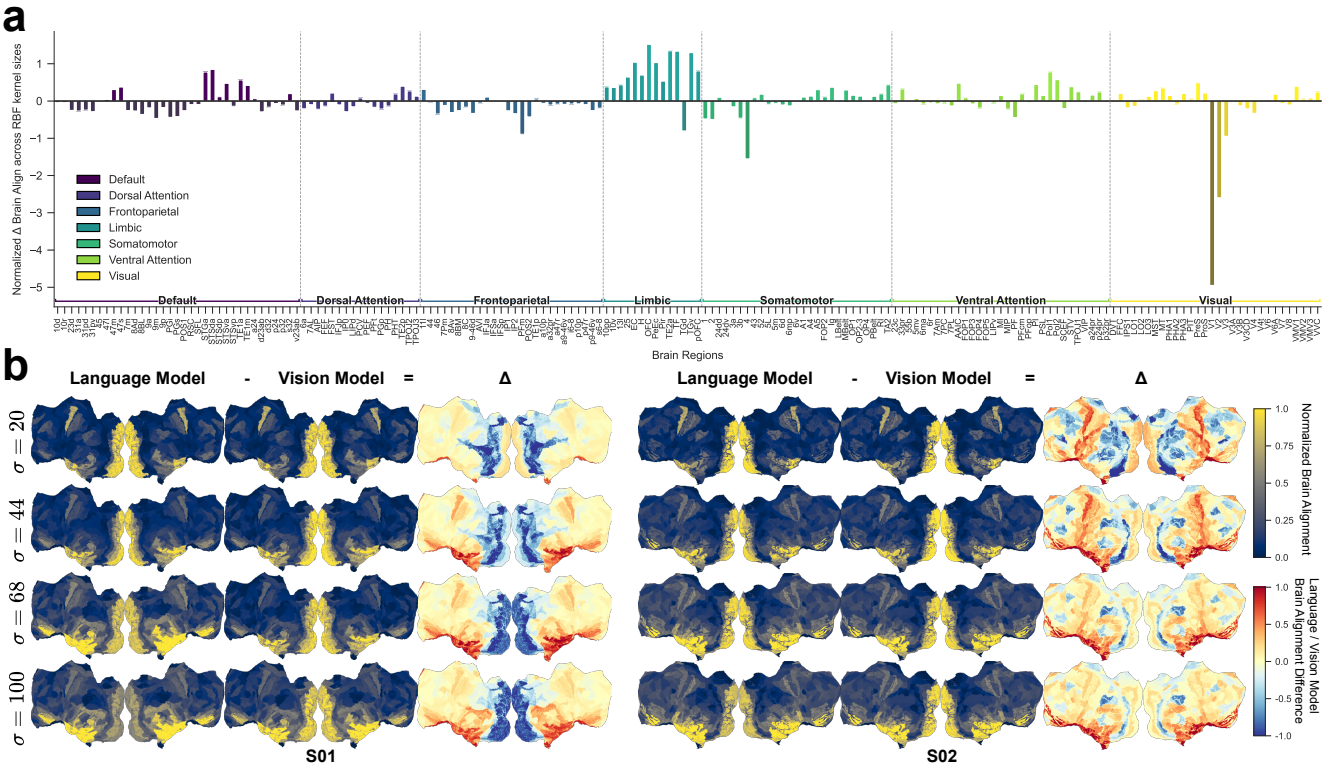

Figure 4. Brain alignment patterns across representational scales. a, Normalized brain alignment changes across different brain regions as kernel size increases from 28 to 68. Positive values indicate regions with increased relative alignment at larger representational scales, while negative values show decreased relative alignment. **b**, Cortical surface visualizations showing alignment patterns between brain activity and both language models (left columns), vision models (middle columns), and their difference (right columns, Δ) at four different kernel sizes. As kernel size increases, a posterior-to-anterior gradient emerges, with anterior regions showing relatively stronger alignment at larger scales, particularly for language models in frontal and temporal areas.

across all brain regions, suggesting a fundamental principle in how information is represented at different scales in the brain.

This shift was visualized on the cortical surface in Figure 4b, revealing a gradient from posterior-to-anterior (occipital to frontal lobe) in alignment patterns at different kernel sizes. The cortical visualizations display the normalized brain alignment scores for both language and vision models across multiple kernel sizes. For language models, stronger alignment is visible in temporal and prefrontal regions compared to primary visual areas, particularly at larger kernel sizes. Vision models show similar patterns but with relatively stronger alignment in posterior visual areas. The difference maps (Δ) highlight regions where language and vision models differ in their brain alignment, with red indicating stronger language model alignment and blue indicating stronger vision model alignment.

As kernel size increases, this posterior-to-anterior gradient becomes more pronounced across all subjects, with anterior regions showing relatively stronger alignment at larger scales. This differentiation is particularly evident for language models in frontal and temporal lobes at the largest kernel sizes. Visualizations for all four subjects and additional kernel size settings are provided in the supplementary materials.

These findings reveal a fundamental organizational principle shared by biological and artificial systems: as information flows through processing pathways, representations transition from local, feature-specific encodings—captured by smaller-scale analyses in primary sensory regions—to distributed, context-dependent structures better captured at larger scales in higher association areas. The stronger scale-dependent modulation observed in language models further suggests that language processing operates at inherently broader representational scales than vision, mirroring hierarchical information processing patterns observed in the brain.

Mapping the Landscape of AI Models via Brain Alignment

Finally, we examined how different models relate to each other when characterized by their brain alignment profiles. Using the full brain alignment pattern consisting of alignment scores across 180 cortical regions, 8 normalized depth levels, and

Figure 5. Taxonomic organization of AI models based on brain alignment patterns. Radial dendrogram showing hierarchical clustering of models based on their brain alignment vectors (180 brain regions $\times$ 8 normalized depth layers $\times$ 4 subjects). Each node represents a model, with node color indicating architecture class (CNN, Transformer, Hybrid), shape indicating modality (circles for language models, stars for vision models), and size representing the alignment score.

all 4 subjects as a high-dimensional feature vector for each model, we constructed a hierarchical clustering tree to visualize representational similarity. The result is shown in Figure 5, where proximity reflects similarity in brain alignment patterns.

Architecture emerges as the dominant organizing principle in the dendrogram. CNNs form two distinct clusters that reflect a generational and structural divide. One smaller cluster contains earlier CNN architectures such as VGG[48] and Xception[49], which lack residual connections and rely on straightforward feedforward stacking of convolutional layers. In contrast, a larger

and more internally coherent cluster includes modern CNNs like ResNet[50], MobileNet[51], and EfficientNet[52], which employ residual or shortcut connections. These architectural innovations appear to yield more brain-aligned representations. The tight grouping of modern CNNs suggests that residual structures impose a stronger inductive bias toward biologically plausible representations, while the earlier models diverge both in form and function.

Transformer-based models occupy a broader and more internally diverse region of the tree. Within this space, language models form several well-defined subclusters that often correspond to model families such as LLaMA[5] or Qwen[6], suggesting that architectural lineage and fine-tuning methodology significantly shape alignment patterns. Language models tend to cluster separately from vision models, even among transformers, indicating a modality-driven divergence in representational geometry at the model level.

Hybrid architectures, which integrate convolutional and attention-based components (e.g., ConvNeXt[53], CoAtNet[54], FocalNet[55]), display a more scattered distribution across the dendrogram. Hybrid architectures display diverse alignment profiles, distributing across the taxonomic space based on their relative balance of convolutional and attentional components. This dispersion reflects their architectural flexibility and suggests that hybrid models do not converge toward a single representational style. Instead, they form a continuum of brain alignment profiles.

This alignment-driven taxonomy offers a biologically grounded organizational view of the model landscape. It complements traditional task benchmarks by revealing deeper commonalities and distinctions in representational structure. These results reinforce the broader theme of our study: representational convergence between artificial and biological systems is structured and quantifiable, and emerges in systematic ways across model families, architectures, and modalities.

Discussion

Our findings provide compelling evidence for convergent evolution between artificial and biological intelligence systems. Despite fundamentally different origins, architectures, and learning mechanisms, AI models spontaneously develop representations that progressively align with human brain activity patterns as they optimize for task performance. This emergent alignment occurs without explicit neurobiological constraints, suggesting that certain computational solutions may be universal for intelligent information processing, transcending specific physical implementations[16,19,56].

Performance Alignment and Convergent Evolution The robust correlation between model performance and brain alignment across both language and vision domains represents perhaps the most striking evidence for convergent evolution. The logarithmic relationship observed in both modalities suggests a principled trajectory of diminishing returns, where initial improvements yield substantial alignment gains followed by more modest increases. This pattern aligns with theoretical perspectives suggesting that as complex systems approach optimal solutions for information processing challenges, they naturally converge toward similar computational strategies despite different evolutionary pathways[9,57]. The substantially stronger correlation observed in language models compared to vision models ($r = 0.89$ vs. $r = 0.53$) suggests that language processing may impose more stringent computational constraints that channel solutions toward brain-like implementations[10,58]. This modality difference may reflect the heightened abstraction and contextual integration demands of language compared to visual perception[59,60].

Hierarchical and Multi-scale Organization Layer-wise and multi-scale alignment analyses revealed systematic gradients of representational organization. Vision models show a gradual increase in brain alignment with depth, paralleling the hierarchical progression from low-level to high-level visual processing in the brain[13,61]. Language models exhibit stronger and more distributed alignment, corresponding to association areas involved in abstract semantic processing[62,63]. Importantly, deeper vision model layers begin to resemble language models, suggesting cross-modal convergence at higher abstraction levels[64,65]. Multi-scale analysis further revealed a posterior-to-anterior shift in alignment across brain regions, consistent with a transition from localized, feature-specific representations to distributed, integrative ones[66,67]. The stronger modulation in language models supports the notion that language operates on inherently broader representational scales[68,69].

Temporal Dynamics and Architectural Inductive Biases Longitudinal analysis revealed that brain alignment consistently precedes improvements in task performance, suggesting that brain-like representations may function as computational stepping stones rather than mere by-products of training[70,71]. This phenomenon was observed across architectures and scales, reinforcing the idea that convergent evolution arises as models optimize for performance. Taxonomic clustering further showed that architectural inductive biases strongly shape alignment profiles. CNNs and transformers formed distinct clusters, while hybrid models exhibited a continuum of profiles. Models with residual connections and attention mechanisms exhibited higher brain alignment scores in our analysis, suggesting these architectural elements may promote representational structures that more closely resemble biological systems[72–74].

Limitations and Implications Our study has several limitations. First, we rely on static fMRI data with limited temporal resolution, which may overlook fine-grained dynamics. Second, the NSD dataset centers on passive visual processing, limiting

insight into language-specific or abstract reasoning processes. Third, benchmark contamination may inflate performance estimates for large models trained on internet-scale corpora[75,76]. Future studies could integrate richer behavioral datasets and temporally resolved neural recordings to address these issues.

These findings have broad implications. For neuroscience, they suggest that AI models optimized for general tasks may spontaneously recapitulate brain-like representations, supporting the idea that cortical organization reflects efficient computational solutions[19,77]. For AI, they imply that brain-aligned architectures or objectives could guide model development and improve interpretability[18,78]. Ultimately, understanding how and why alignment emerges may bridge the gap between natural and artificial intelligence.

Methods

To systematically investigate brain-AI alignment, we developed a comprehensive analytical framework combining neuroimaging and machine learning techniques. This approach quantifies similarities between neural activity patterns in human brains and internal representations in AI models using Centered Kernel Alignment (CKA)[24], a robust similarity metric that accounts for different dimensionalities across systems. We analyzed fMRI responses from human subjects viewing natural images (Natural Scenes Dataset)[20] alongside activations from 630 AI models (spanning vision and language domains) processing the same stimuli. By varying measurement parameters and examining alignment across different brain regions, model architectures, and training stages, we uncovered systematic patterns of representational convergence between biological and artificial systems. Statistical robustness was ensured through extensive validation across multiple subjects, metrics, and analytical parameters.

Data Acquisition and Preprocessing

fMRI Data For our analysis, we utilized the Natural Scenes Dataset (NSD)[20], a large-scale fMRI dataset capturing neural responses to natural images. From the eight subjects in the dataset, we selected four (S01, S02, S05, S07) who had complete data trajectories. Each subject's data included responses to 24,980 training stimuli (unique per subject) and 2,770 testing stimuli (shared across subjects). To maximize data volume while ensuring consistent analysis conditions, we focused on the training set for our primary analyses.

We extracted functional responses from the preprocessed `func1pt8mm` data following the standard NSD preprocessing pipeline, which includes motion correction, distortion correction, and hemodynamic response modeling to extract beta weights representing neural responses to each stimulus. All analyses were conducted in subject-specific functional space to preserve individual neural response patterns.

Stimulus Materials The NSD stimuli consist of images from the COCO dataset[21], which provides paired image-caption annotations. For each image, we extracted the first available caption to serve as the textual input for language models. This approach enabled aligned multimodal analyses across vision and language domains using identical stimuli, allowing for direct comparison of neural representations across modalities.

Neural Network Models and Representation Extraction

Language Models We analyzed 36 contemporary language models spanning diverse architectures and parameter scales (500M to 72B parameters). Our selection included major model families: Qwen[6], Llama[5], Gemma[22], Mistral[79], and Phi[80], among others[81]. To explore the effects of different fine-tuning approaches, we included specialized variants such as SimPO[32] and Tulu[31].

For consistent representation extraction, we implemented all models using the HuggingFace Transformers library[82] with default configurations. Each model processed image captions using its standard chat template. For example, with Llama-3 models, the template followed this structure:

```
<|start_header_id|>user<|end_header_id|>
Consider the image described by the following sentences:
    '{caption}'
<|start_header_id|>assistant<|end_header_id|>
```

From each model, we extracted hidden state representations from the output of each transformer block, focusing on the last token representation, which is a standard approach for capturing sentence-level semantics[8,83,84]. For performance evaluation, we used LLM Leaderboard 2[29], a composite benchmark aggregating performance across multiple tasks including BBH (Big-Bench Hard)[34], MMLU-PRO[38], IFEval[33], GPQA[36], MATH[35], and MUSR[37], covering various reasoning, knowledge, and language understanding dimensions. We also incorporated Chatbot Arena rankings[39][39], which offers complementary human evaluations through a pairwise preference system rather than automated metrics.

Vision Models We analyzed 594 vision models spanning three architectural categories: Convolutional Neural Networks (CNNs), Transformers, and Hybrid models. All models were sourced from the TIMM library[85] with pre-trained weights, encompassing parameter ranges from 1.33M to 1.014B. Due to the heterogeneous structure of vision architectures, we implemented an automated feature extraction pipeline to capture hidden states from each network block across all models, extracting outputs from each individual feature extraction component throughout the network hierarchy.

For each image stimulus, we applied the standard preprocessing procedure specified for each model (typically normalizing according to ImageNet statistics). Model performance was evaluated using multiple benchmarks, including ImageNet-1K top-1 accuracy[30], ImageNet-A[40], ImageNet-R[41], and other variant datasets as detailed in Table 1.

Brain Parcellation and Anatomical Framework

We employed the HCP_MMP1 parcellation[25], which divides the cortex into 180 anatomically and functionally distinct regions. To analyze functional network properties, we mapped these regions to the Yeo-7 network taxonomy[26], which groups cortical areas into seven functional networks: Visual, Somatomotor, Dorsal Attention, Ventral Attention, Limbic, Frontoparietal, and Default. This mapping followed the correspondence established by Byrge and Kennedy[86], enabling analysis of representational similarities at both region-specific and network levels.

Representational Similarity Analysis

Centered Kernel Alignment To quantify representational similarity between AI model activations and brain activity patterns, we employed CKA, a robust similarity index that accounts for different dimensionalities and is invariant to orthogonal transformations. The CKA between two representation matrices X and Y is defined as:

$$\mathrm{CKA}(K,L) = \frac{\mathrm{HSIC}(K,L)}{\sqrt{\mathrm{HSIC}(K,K) \times \mathrm{HSIC}(L,L)}}, \tag{1}$$

where K and L are kernel matrices derived from X and Y, and HSIC is the Hilbert-Schmidt Independence Criterion. For our analyses, we utilized the Radial Basis Function (RBF) kernel:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \tag{2}$$

where σ controls the kernel width. For computational efficiency with high-dimensional representations exceeding 4096 dimensions, we applied Principal Component Analysis (PCA) to reduce dimensionality while preserving at least 95% of variance. All CKA computations were implemented in PyTorch and executed on GPUs with parallelization across models and brain regions.

We performed analyses across multiple kernel sizes (20, 28, 36, 44, 52, 60, 68, 72, 84, 92, 100) to capture representational similarities at different scales, from local feature-specific patterns to global representational structures. This multi-scale approach enabled detection of alignment patterns that might be specific to particular representational granularities. Despite observing scale-dependent patterns as reported in the Results section, we found high consistency in alignment patterns across different kernel sizes, confirming the robustness of our observations.

To further validate the stability of our similarity metrics, we conducted a comprehensive consistency analysis across different kernel sizes and alternative k-nearest neighbor (KNN) based similarity measures (Figure 6). Using Spearman rank correlation, we evaluated whether the relative rankings of model similarities remained consistent despite variations in the similarity computation method. This analysis confirmed strong rank consistency across different parameter settings, supporting the reliability of our representational similarity framework regardless of specific kernel size choices.

For brain-region specific analyses, we calculated alignment scores for each model by taking the average of the top 5% highest-aligning brain regions at the normalized model depth of 6/8. This approach highlighted the brain regions most strongly corresponding to model representations, providing a more sensitive measure of alignment than averaging across all regions.

Longitudinal Training Analysis

To investigate how brain alignment evolves during model training, we leveraged two model families with available training checkpoints:

- For language models, we analyzed the Pythia family[27], which provides numerous checkpoints throughout training for models ranging from 70M to 2.8B parameters. These checkpoints enabled tracking of representational evolution across training trajectories.

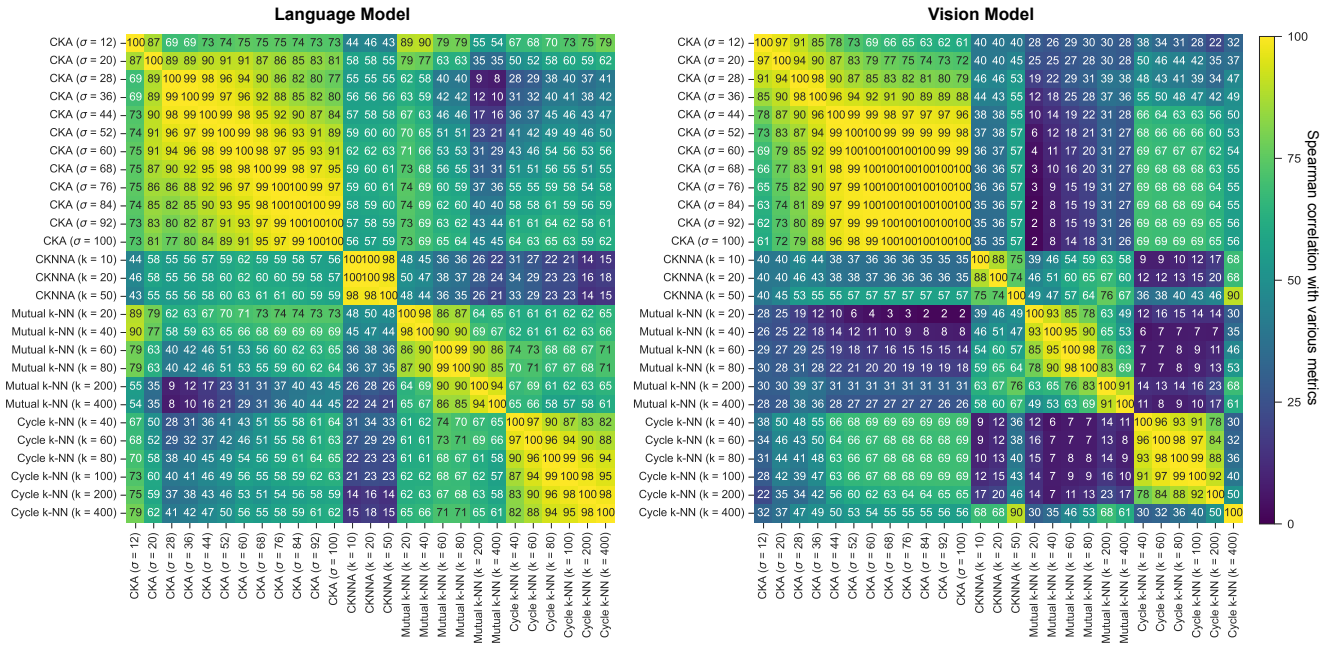

Figure 6. Consistency analysis across different similarity metrics and parameter settings. Heatmaps showing Spearman rank correlations between different similarity measurement approaches for language models and vision models. The analysis compares CKA with varying kernel sizes, k-nearest neighbor (KNN) variants with different k values. The consistently high correlation values indicate strong agreement in how different metrics rank the similarity relationships, demonstrating the robustness of our representational alignment framework across methodological variations.

- For vision models, we utilized the MixNet family[28], training multiple variants (S/M/L/XL) from scratch on ImageNet-1K[23] using standard TIMM[85] configurations. Training checkpoints were saved at logarithmically spaced intervals to capture both early and late training dynamics.

For language models, we tracked performance on LAMBADA[46] (language modeling) and PIQA[47] (physical common sense reasoning) benchmarks across training checkpoints. For vision models, we monitored ImageNet top-1 and top-5 accuracy. At each checkpoint, we calculated brain alignment scores using the CKA methodology described above, enabling temporal correlation analysis between brain alignment and task performance. Pearson correlation coefficients were computed to quantify the relationship between alignment and performance trajectories throughout training.

Multi-scale Analysis Methods

Regional Sensitivity to Kernel Size To analyze how different brain regions respond to variations in representational scale, we compared alignment patterns across different kernel sizes. For each brain region, we calculated normalized alignment scores by first averaging across all models and then computing the difference (Δ) in alignment between larger kernel sizes (68) and smaller kernel sizes (28). Positive Δ values indicate regions with relatively stronger alignment at larger scales (capturing more global representational properties), while negative values indicate regions with stronger alignment at smaller scales (capturing more local features).

To validate the consistency of these patterns, we calculated the sign consistency of normalized alignment changes across subjects, finding a high consistency value of 0.83, which confirms the robustness of the observed scale-dependent regional preferences across individuals.

Model Taxonomy Construction

We constructed a hierarchical organization of models based on their brain alignment profiles using agglomerative clustering. For each model, we created a feature vector comprising alignment scores across all 180 brain regions and 8 normalized depth levels for all 4 subjects, resulting in a $180 \times 8 \times 4$ dimensional representation. We computed pairwise distances using cosine similarity and applied Ward's linkage method[87] to generate the hierarchical clustering structure. The resulting dendrogram was visualized as a radial tree diagram with node attributes (size, color, shape) representing brain alignment score, architectural class, and modality, respectively.

Statistical Analysis Framework

Correlation Analysis We assessed the relationship between model performance and brain alignment using both Pearson and Spearman correlation coefficients, capturing linear and monotonic relationships respectively. For each correlation, we calculated 95% confidence intervals via bootstrap sampling with 1,000 iterations. To account for multiple comparisons across different performance metrics and model types, we applied False Discovery Rate (FDR) correction to p-values.

Regression Models To characterize the relationship between model performance and brain alignment, we fitted both linear and logarithmic regression models. Model selection was based on Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and R^2 values. For the language model performance-alignment relationship, a logarithmic model ($R^2 = 0.80$, AIC = 117.6, BIC = 120.8) outperformed a linear model ($R^2 = 0.78$, AIC = 120.7, BIC = 123.9), indicating a diminishing returns pattern. Similarly, for vision models, a logarithmic fit ($R^2 = 0.33$, AIC = -3810.2, BIC = -3801.4) provided better characterization than a linear model ($R^2 = 0.28$, AIC = -3772.1, BIC = -3763.3).

Table 2. Inter-subject consistency for language models. Pearson correlation (r) values between brain alignment patterns across subjects for language models.

	S01	S02	S05	S07
S01	1.000	0.997	0.999	0.997
S02	0.997	1.000	0.999	1.000
S05	0.999	0.999	1.000	0.999
S07	0.997	1.000	0.999	1.000

Table 3. Inter-subject consistency for vision models. Pearson correlation (r) values between brain alignment patterns across subjects for vision models.

	S01	S02	S05	S07
S01	1.000	0.970	0.981	0.970
S02	0.970	1.000	0.998	0.999
S05	0.980	0.998	1.000	0.998
S07	0.970	0.999	0.998	1.00

Inter-subject Consistency Analysis To ensure the robustness of our findings, we conducted a comprehensive analysis to determine whether the relationship between brain alignment scores and model performance was consistent across different individuals. We calculated pairwise Pearson correlations between these brain-performance relationships for all four subjects (S01, S02, S05, S07), as shown in Tables 2 and 3. These tables demonstrate the consistency of how brain alignment relates to model performance across different individuals.

For language models, we observed remarkably high inter-subject correlations, with values ranging from 0.997 to 1.000. This near-perfect consistency indicates that the pattern of which brain regions align with language model representations is highly preserved across different individuals. For vision models, we also found extremely strong consistency, with correlations ranging from 0.970 to 0.999.

These extraordinarily high inter-subject correlations for both modalities provide compelling evidence that the regional patterns of brain-AI alignment we observed reflect fundamental organizational principles of neural information processing universally shared across human brains rather than idiosyncratic individual variations.

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436–444 (2015).
2. Silver, D. *et al.* Mastering the game of go with deep neural networks and tree search. *nature* **529**, 484–489 (2016).
3. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).
4. Jones, C. R. & Bergen, B. K. Large language models pass the turing test. *arXiv preprint arXiv:2503.23674* (2025).
5. Grattafiori, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
6. Yang, A. *et al.* Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115* (2024).
7. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
8. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
9. Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. national academy sciences* **111**, 8619–8624 (2014).
10. Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci.* **118**, e2105646118 (2021).

11. Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J. & Kriegeskorte, N. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *J. cognitive neuroscience* **33**, 2044–2064 (2021).

12. Zhuang, C., Kubilius, J., Hartmann, M. J. & Yamins, D. L. Toward goal-driven neural network models for the rodent whisker-trigeminal system. *Adv. Neural Inf. Process. Syst.* **30** (2017).

13. Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. neuroscience* **19**, 356–365 (2016).

14. Nilsson, D.-E. Eye evolution and its functional basis. *Vis. neuroscience* **30**, 5–20 (2013).

15. Ogura, A., Ikeo, K. & Gojobori, T. Comparative analysis of gene expression for convergent evolution of camera eye between octopus and human. *Genome research* **14**, 1555–1561 (2004).

16. Zador, A. M. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat. communications* **10**, 3770 (2019).

17. Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. Backpropagation and the brain. *Nat. Rev. Neurosci.* **21**, 335–346 (2020).

18. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258 (2017).

19. Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nat. neuroscience* **22**, 1761–1770 (2019).

20. Allen, E. J. *et al.* A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. neuroscience* **25**, 116–126 (2022).

21. Lin, T.-Y. *et al.* Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755 (Springer, 2014).

22. Team, G. *et al.* Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024).

23. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. neural information processing systems* **25** (2012).

24. Kornblith, S., Norouzi, M., Lee, H. & Hinton, G. Similarity of neural network representations revisited. In *International conference on machine learning*, 3519–3529 (PMLR, 2019).

25. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).

26. Yeo, B. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. neurophysiology* (2011).

27. Biderman, S. *et al.* Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2397–2430 (PMLR, 2023).

28. Tan, M. & Le, Q. V. Mixconv: Mixed depthwise convolutional kernels. In *BMVC* (2019).

29. Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K. & Wolf, T. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard (2024).

30. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).

31. Lambert, N. *et al.* T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124* (2024).

32. Meng, Y., Xia, M. & Chen, D. Simpo: Simple preference optimization with a reference-free reward. *Adv. Neural Inf. Process. Syst.* **37**, 124198–124235 (2024).

33. Zhou, J. *et al.* Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911* (2023).

34. Suzgun, M. *et al.* Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, 13003–13051 (2023).

35. Hendrycks, D. *et al.* Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2024).

36. Rein, D. *et al.* Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling* (2024).

37. Sprague, Z., Ye, X., Bostrom, K., Chaudhuri, S. & Durrett, G. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *ICLR* (2024).

38. Wang, Y. *et al.* Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2024).

39. Chiang, W.-L. *et al.* Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning* (2024).

40. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J. & Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15262–15271 (2021).

41. Hendrycks, D. *et al.* The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8340–8349 (2021).

42. Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X. & Oord, A. v. d. Are we done with imagenet? *arXiv preprint arXiv:2006.07159* (2020).

43. Recht, B., Roelofs, R., Schmidt, L. & Shankar, V. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, 5389–5400 (PMLR, 2019).

44. Wang, H., Ge, S., Lipton, Z. & Xing, E. P. Learning robust global representations by penalizing local predictive power. *Adv. neural information processing systems* **32** (2019).

45. McInnes, L., Healy, J., Saul, N. & Großberger, L. Umap: Uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).

46. Paperno, D. *et al.* The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1525–1534 (2016).

47. Bisk, Y., Zellers, R., Gao, J., Choi, Y. *et al.* Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 7432–7439 (2020).

48. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

49. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258 (2017).

50. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

51. Koonce, B. & Koonce, B. Mobilenetv3. *Convolutional Neural Networks with Swift for Tensorflow: Image Recognit. Dataset Categ.* 125–144 (2021).

52. Koonce, B. Efficientnet. In *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization*, 109–123 (Springer, 2021).

53. Liu, Z. *et al.* A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986 (2022).

54. Dai, Z., Liu, H., Le, Q. V. & Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Adv. neural information processing systems* **34**, 3965–3977 (2021).

55. Yang, J., Li, C., Dai, X. & Gao, J. Focal modulation networks. *Adv. Neural Inf. Process. Syst.* **35**, 4203–4217 (2022).

56. Hasson, U., Nastase, S. A. & Goldstein, A. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron* **105**, 416–434 (2020).

57. Marblestone, A. H., Wayne, G. & Kording, K. P. Toward an integration of deep learning and neuroscience. *Front. computational neuroscience* **10**, 94 (2016).

58. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. biology* **5**, 134 (2022).

59. Fedorenko, E. & Thompson-Schill, S. L. Reworking the language network. *Trends cognitive sciences* **18**, 120–126 (2014).

60. Pereira, F. *et al.* Toward a universal decoder of linguistic meaning from brain activation. *Nat. communications* **9**, 963 (2018).

61. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. reports* **6**, 27755 (2016).

62. Goldstein, A. *et al.* Shared computational principles for language processing in humans and deep language models. *Nat. neuroscience* **25**, 369–380 (2022).

63. Dobs, K., Martinez, J., Kell, A. J. & Kanwisher, N. Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci. advances* **8**, eabl8913 (2022).

64. Binder, J. R. *et al.* Toward a brain-based componential semantic representation. *Cogn. neuropsychology* **33**, 130–174 (2016).

65. Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).

66. Margulies, D. S. *et al.* Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl. Acad. Sci.* **113**, 12574–12579 (2016).

67. Huntenburg, J. M., Bazin, P.-L. & Margulies, D. S. Large-scale gradients in human cortical organization. *Trends cognitive sciences* **22**, 21–31 (2018).

68. Fedorenko, E. & Varley, R. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Annals New York Acad. Sci.* **1369**, 132–153 (2016).

69. Hagoort, P. The neurobiology of language beyond single-word processing. *Science* **366**, 55–58 (2019).

70. Saxe, A. M., McClelland, J. L. & Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proc. Natl. Acad. Sci.* **116**, 11537–11546 (2019).

71. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. review vision science* **1**, 417–446 (2015).

72. Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nat. neuroscience* **21**, 1148–1160 (2018).

73. Kietzmann, T. C. *et al.* Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci.* **116**, 21854–21863 (2019).

74. Kubilius, J. *et al.* Brain-like object recognition with high-performing shallow recurrent anns. *Adv. neural information processing systems* **32** (2019).

75. Dodge, J. *et al.* Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758* (2021).

76. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623 (2021).

77. Kriegeskorte, N. & Douglas, P. K. Interpreting encoding and decoding models. *Curr. opinion neurobiology* **55**, 167–179 (2019).

78. Lindsay, G. W. Convolutional neural networks as a model of the visual system: Past, present, and future. *J. cognitive neuroscience* **33**, 2017–2031 (2021).

79. Jiang, A. Q. *et al.* Mistral 7B, DOI: 10.48550/arXiv.2310.06825. 2310.06825.

80. Abdin, M. *et al.* Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* (2024).

81. Team, I. Internlm: A multilingual language model with progressively enhanced capabilities (2023).

82. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (Association for Computational Linguistics, Online, 2020).

83. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186 (2019).

84. Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

85. Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, DOI: 10.5281/zenodo.4414861 (2019).

86. Byrge, L. & Kennedy, D. P. High-accuracy individual identification using a "thin slice" of the functional connectome. *Netw. Neurosci.* **3**, 363–383 (2019).

87. Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *J. Am. statistical association* **58**, 236–244 (1963).

Author contributions statement

G.S. performed the primary analyses, and implemented the computational framework. D.Z. contributed to the conceptualization of the research idea. G.S., D.Z., and Q.Z. jointly designed the experiments, established the analytical pipeline, and developed the methodological approach. Y.D. assisted with data processing and analysis. Y.Z. supervised the project and provided funding support. G.S. wrote the original draft. All authors contributed to reviewing and editing the manuscript. Y.Z. provided critical revisions of the manuscript and approved the final version.

Data Availability

All brain imaging data used in this study is from the Natural Scenes Dataset (NSD) (https://naturalscenesdataset.org/)[20], a large-scale fMRI dataset containing neural responses to natural images. The image stimuli and captions were sourced from the COCO dataset (https://cocodataset.org/)[21]. For language model analyses, we used publicly available pre-trained models from the HuggingFace Transformers library (https://huggingface.co/models)[82], including models from the Qwen[6], Llama[5], Gemma[22], Mistral[79], Phi[80] families and others[27,31,32,81]. For vision model analyses, we used pre-trained models from the TIMM library (https://github.com/huggingface/pytorch-image-models)[85]. Model performance metrics were obtained from LLM Leaderboard 2 (https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard)[29] for language models and standard ImageNet[23] benchmarks for vision models. For longitudinal analyses, we used checkpoints from the Pythia model family (https://github.com/EleutherAI/pythia)[27] and trained MixNet models[28] with saved checkpoints.

Code Availability

All code used for data processing, model representation extraction, alignment computation, and statistical analyses is available on https://github.com/FloyedShen/BrainAlign.