Convergent Evolution across Modalities, Scales and Training Trajectories: Evidence for Human Brain–AI Alignment

Guobin Shen · **Dongcheng Zhao** · **Yiting Dong** · **Qian Zhang** · **Yi Zeng**

Institute of Automation, Chinese Academy of Sciences (CASIA)

ICML International Conference On Machine Learning

Motivation & Approach

As AI matches human performance, do its **internal representations converge toward the brain**? Prior work is confined to single modalities, narrow brain regions, and static training checkpoints.

We frame this as **convergent evolution**: systems under shared pressure may reach similar solutions across modalities, scales, and time.

What we do

- **630 AI models** — 36 LLMs (0.5–72B) and 594 vision models (1.33M–1B), spanning diverse architectures.
- Compared against **NSD fMRI** from 4 subjects viewing natural images with matched captions.
- **CKA** alignment over layers × 180 cortical regions × kernels → **>60M** measurements.

Key findings

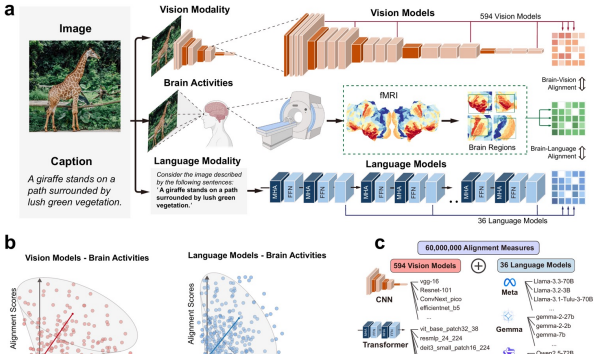
1 Performance–alignment correlation
Stronger performance predicts closer alignment (language **r=0.89**, vision **r=0.53**).

2 Alignment emerges early in training
Early alignment Granger-predicts performance (**9/10 vs 3/10**).

3 Modality-specific anatomy
Vision → visual hierarchy; language → limbic / default-mode.

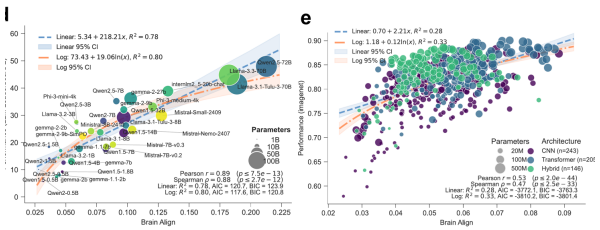
4 Posterior→anterior scale gradient
Larger kernels shift sensory → association cortex.

Framework & Performance



630 models × NSD fMRI; vision sees images, language sees captions of the same scenes.

Performance–alignment

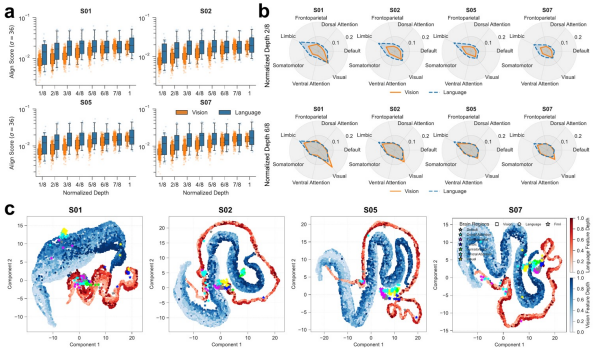


Language r = 0.89 **Vision r = 0.53** (log fit, FDR-corrected)

Holds after controlling for model size (partial r = 0.54 / 0.31).

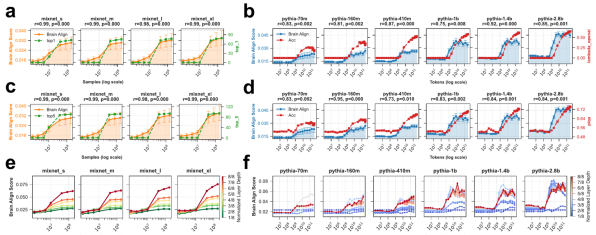
Capability, not parameter count, is what tracks brain alignment.

Hierarchy & Dynamics



Vision rises with depth; language peaks mid-network. Deeper vision layers approach the language manifold (Wasserstein ↓, ρ=−1).

Alignment precedes performance

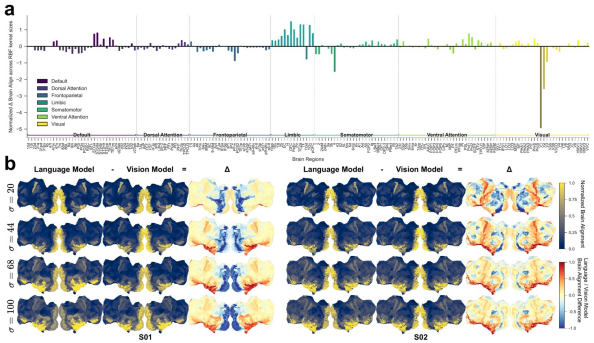


Across training, **brain alignment rises before task accuracy**.

Granger align→perf: **9/10** perf→align: **3/10**

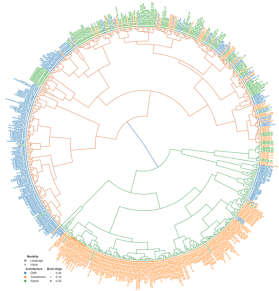
Brain-like structure forms early and scaffolds later task gains.

Scale, Taxonomy & Robustness



Small kernels → **posterior sensory cortex** (V1 **Δ=−5.3**); large kernels → **anterior association cortex** (OFC **Δ=+1.6**; sign consistency 0.83).

Model taxonomy & robustness



Profiles cluster by **architecture**, then modality.

Mantel vs benchmark ≈ 0: brain adds structure benchmarks miss.

Top LLMs reach the **noise ceiling** in 88% of regions.

Optimizing for task performance drives AI toward human-like computation.