LANSDOWNE
PARTNERS

Large Language Models Vs The Brain

By Professor Justin Stebbing

Executive Summary

Both the human brain and large language models ("LLMs") process images, but their forms of perception, while sharing parallels, are distinct in origin, structure, and implication. Philosophers have debated the nature of human perception for centuries. David Hume proposed that all mental life consists of "perceptions," which he divided into "impressions" (vivid, forceful sensations or emotions) and "ideas" (weaker copies of impressions). According to Hume, even the most abstract concepts are built up from elementary sensory data, yet crucially, these data are not "pixels" per se, but rather forceful, complex experiences shaped by the senses. The mind, for Hume, does not directly access the world; instead, it manipulates these impressions and ideas to construct understanding. Immanuel Kant, on the other hand, advanced a radically different formulation: perception is not a passive reception of data, but an active structuring of sensory input according to pre-existing categories and intuitions (notably, space and time). For Kant, what is perceived is always "as it appears", structured by the mind's faculties, not experienced as the "thing in itself". Thus, the "image" seen by the brain is never raw input but an already-interpreted phenomenon.

AI scientists are now joining the debate and its implications for model design. Yann LeCun argues that human-level (or animal) intelligence emerges from learning "world models" grounded in experience, but this is a capacity AI lacks in its current form. LeCun's goal is not to make machines see raw pixels, but to enable them to build internal representations of the world: structures that support reasoning, prediction, and planning. These internal representations may be trained on pixels, but their very utility arises only once low-level sensory data is abstracted into meaningful patterns. Yoshua Bengio (currently making headlines with over a million citations on Google Scholar[1]) has emphasized that intelligence in deep learning arises not from handling raw data brutishly, but from learning layered, structured representations, features and concepts that allow for generalization and abstraction. Tools like attention mechanisms, hierarchical features, and transfer learning serve a role analogous to the organizing faculties that Kant described: abstracting from data to structured understanding.

Recent studies show that multimodal LLMs (like GPT-Vision) and the human brain exhibit strikingly similar ways of clustering and representing images semantically, even if their building blocks rather obviously differ. In both cases, the process is not a mere "reading of pixels" but involves creating high-dimensional embeddings or distributed patterns of neural activity encoding meaning and relationships. The alignment is found in the representational "geometry": how objects are grouped, compared, and interpreted, not in literal equivalence of pixel-data to neuronal firing. Here, AI models and human brains do not simply "see pixels." Perception in both is a layered, representational process: the brain actively organizes sensory input using a priori categories (Kant), forming impressions and ideas (Hume), while LLMs (and their multimodal descendants) abstract images into latent feature spaces that express meaning rather than raw data (LeCun, Bengio). New data confirms that these representational spaces can align across human and artificial cognition. Thus, to ask if LLMs "see just pixels" is to misunderstand the architectures of perception, past and present, biological and artificial. Both systems conduct sophisticated transformations that underscore the constructed, structured nature of all "vision", never a simple reading of the raw.

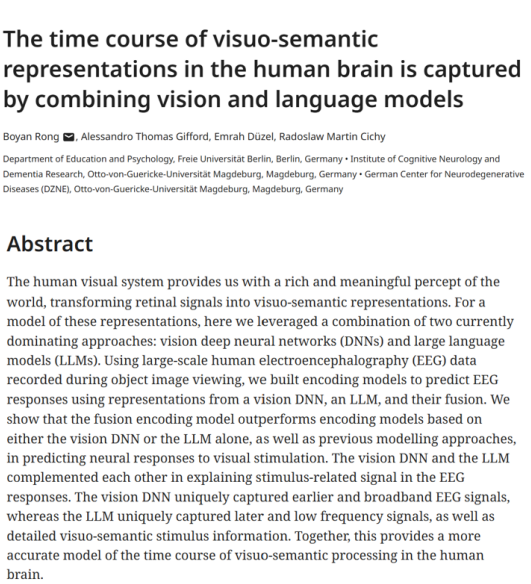
[1] Else, Holly. "AI Pioneer Becomes First Living Scientist to Hit One Million Citations." Nature, November 20, 2025. https://www.nature.com/articles/d41586-025-03681-6.

Introduction

Geoffrey Hinton differentiates between human "type two" intelligence—deliberate, logical reasoning—and the more immediate, intuitive "type one" intelligence central to perception and analogy making. Hinton's GLOM architecture aims to model perception in machines more like humans, through recursive part-whole relationships and analogical pattern mapping. According to Hinton, AI intuition does not depend on processing raw pixels but on extracting and organizing semantic representations, what he refers to as "big vectors", that serve as analogues to human cognition. Douglas Hofstadter, famed for "Gödel, Escher, Bach," introduces the concept of "strange loops" in both consciousness and AI. He suggests that systems capable of abstraction, recursion, and self-reference are those closest to manifesting true understanding and meaning. Hofstadter argues that advanced AI models go beyond pixel-level computation by recursively modelling their own operations and generating multi-layered abstractions that can sometimes "emulate" aspects of consciousness, despite not possessing unified subjectivity.

Most work in deep learning reinforces that perception in machines arises from hierarchical representations, layers of abstraction that transform raw sensory input into structured, meaningful feature spaces. Here, neural nets learn to ignore irrelevant input and focus on salient features, much like the Kantian mind applies categories of understanding to organize sensory data. For example, Bengio's concept of priors in AI echoes Kant's "a priori intuitions," serving as scaffolding for learning and generalization.

This month, an empirical study presents, in my view, compelling evidence that LLMs and vision deep neural networks (DNNs) create high-dimensional embeddings and organizational schemas similar to those observed in human neural activity during perception of images[2]. For example, combining a DNN's vision model with a language model enables better prediction of neural responses to visual stimuli than using either model alone, mirroring the layered, integrative approach of human perception (lower left figure). In fact, LLMs extract semantic relationships across pixels, clustering visually or conceptually similar objects in latent space—analogous to the brain's distributed patterns of activation and semantic clustering:



The time course of visuo-semantic representations in the human brain is captured by combining vision and language models

Boyan Rong ✉, Alessandro Thomas Gifford, Emrah Düzel, Radoslaw Martin Cichy

Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany • Institute of Cognitive Neurology and Dementia Research, Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany • German Center for Neurodegenerative Diseases (DZNE), Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany

Abstract

The human visual system provides us with a rich and meaningful percept of the world, transforming retinal signals into visuo-semantic representations. For a model of these representations, here we leveraged a combination of two currently dominating approaches: vision deep neural networks (DNNs) and large language models (LLMs). Using large-scale human electroencephalography (EEG) data recorded during object image viewing, we built encoding models to predict EEG responses using representations from a vision DNN, an LLM, and their fusion. We show that the fusion encoding model outperforms encoding models based on either the vision DNN or the LLM alone, as well as previous modelling approaches, in predicting neural responses to visual stimulation. The vision DNN and the LLM complemented each other in explaining stimulus-related signal in the EEG responses. The vision DNN uniquely captured earlier and broadband EEG signals, whereas the LLM uniquely captured later and low frequency signals, as well as detailed visuo-semantic stimulus information. Together, this provides a more accurate model of the time course of visuo-semantic processing in the human brain.

High-level visual representations in the human brain are aligned with large language models

Adrien Doerig, Tim C. Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay & Ian Charest ✉

Nature Machine Intelligence **7**, 1220–1234 (2025) | Cite this article

51k Accesses | **13** Citations | **167** Altmetric | Metrics

A preprint version of the article is available at arXiv.

Abstract

The human brain extracts complex information from visual inputs, including objects, their spatial and semantic interrelations, and their interactions with the environment. However, a quantitative approach for studying this information remains elusive. Here we test whether the contextual information encoded in large language models (LLMs) is beneficial for modelling the complex visual information extracted by the brain from natural scenes. We show that LLM embeddings of scene captions successfully characterize brain activity evoked by viewing the natural scenes. This mapping captures selectivities of different brain areas and is sufficiently robust that accurate scene captions can be reconstructed from brain activity. Using carefully controlled model comparisons, we then proceed to show that the accuracy with which LLM representations match brain representations derives from the ability of LLMs to integrate complex information contained in scene captions beyond that conveyed by individual words. Finally, we train deep neural network models to transform image inputs into LLM representations. Remarkably, these networks learn representations that are better aligned with brain representations than a large number of state-of-the-art alternative models, despite being trained on orders-of-magnitude less data. Overall, our results suggest that LLM embeddings of scene captions provide a representational format that accounts for complex information extracted by the brain from visual inputs.

[2] Rong, Boyan, Alessandro Thomas, Gifford Emrah Düzel, and Radoslaw Martin Cichy. "The Time Course of Visuo-Semantic Representations in the Human Brain Is Captured by Combining Vision and Language Models." eLife (reviewed preprint), November 5, 2025. https://elifesciences.org/reviewed-preprints/108915.

The human brain extracts complex information from visual inputs, including objects, their spatial and semantic interrelations and their interactions with the environment. In an impressive paper published this summer (upper right figure), a research team investigated whether the contextual information encoded in large LLMs is beneficial for modelling the complex visual information extracted by the brain from natural scenes[3]. The study demonstrates that LLM embeddings of scene captions accurately characterize brain activity evoked by viewing the natural scenes. This mapping captures the selectivity of different brain areas and is robust enough to allow for the reconstruction of accurate scene captions from brain activity. Through carefully controlled model comparisons, the team shows that the accuracy of LLM representations in matching brain representations derives from the ability of LLMs to integrate complex information contained in scene captions beyond information conveyed by individual words. Additionally, they train DNN models to transform image inputs into LLM representations. Remarkably, these networks learn representations that align more closely with brain activity than many state-of-the-art alternative models, despite being trained on orders-of-magnitude less data. Overall, these results show that LLM embeddings of scene captions provide a representational format that accounts for complex information extracted by the brain from visual inputs.

Both systems also exhibit "fusion" stages: early stages of perception dominated by sensory encoding, followed by higher-level conceptual and linguistic interpretation. Yet, crucial qualitative differences remain: the brain's perception is rooted in bodily, lived context (Merleau-Ponty), refined by judgment (Descartes), and transformed by aspectual seeing (Wittgenstein). LLMs, while able to produce human-like outputs, fundamentally lack embodiment, intentionality and direct experiential meaning. Other comparisons can be made here too[4]:

Feature	Human Brain (Philosophers)	AI Systems (Hinton, Hofstadter, etc.)
Raw Input	Sensation, embodied experience	Pixel arrays, sensory tensors
Aspectual/Contextualization	Seeing-as, aspect change	Attention, hierarchical feature extraction
Judgment/Abstraction	Judgment, reason, intuition	Analogical reasoning, vector embeddings
Bodily Foundation	Primacy of perception, embodiment	Synthetic: lacks lived context
Self-reference/Meaning	Strange loops, intentionality	Recursion, strange loops, abstraction

The table above using representational similarity analysis (RSA), encoding models, linear decoding, and artificial neural network (ANN) modelling, provides evidence that the human visual system, especially in higher-level regions, develops representations that closely match those found in the embeddings of LLMs trained on descriptive captions of visual scenes. This is particularly noteworthy because LLMs lack direct visual experience; instead, the statistical patterns and contextual knowledge they acquire through extensive language training reflect the complexities extracted by the brain's sensory systems when processing visual inputs.

LLMs can thus be used leveraged to enhance the representations used in visual models. As well as neuroscientific work highlighting similarities between linguistic and visual representations in the brain and showing that linguistic information improves the ability of crossmodal ANNs to predict brain activities, the approach is complementary to, rather than competitive with, existing work on high-level visual features—such as object- and scene-category extraction, linguistic aspects, and the statistical occurrence of objects and their locations. While past research has typically analysed these features in isolation, LLM embeddings offer a unified, quantitative framework that encompasses category information and models diverse elements of visual processing together. These findings suggest that LLM-trained ANNs can capture category information alongside other feature-based aspects, indicating the

[3] Doerig, Adrien, Tim C. Kietzmann, and Nikolaus Kriegeskorte. "High-Level Visual Representations in the Human Brain Are Aligned with Large Language Models." Nature Machine Intelligence, published online August 7, 2025. https://www.nature.com/articles/s42256-025-01072-0.

[4] Author's own work.

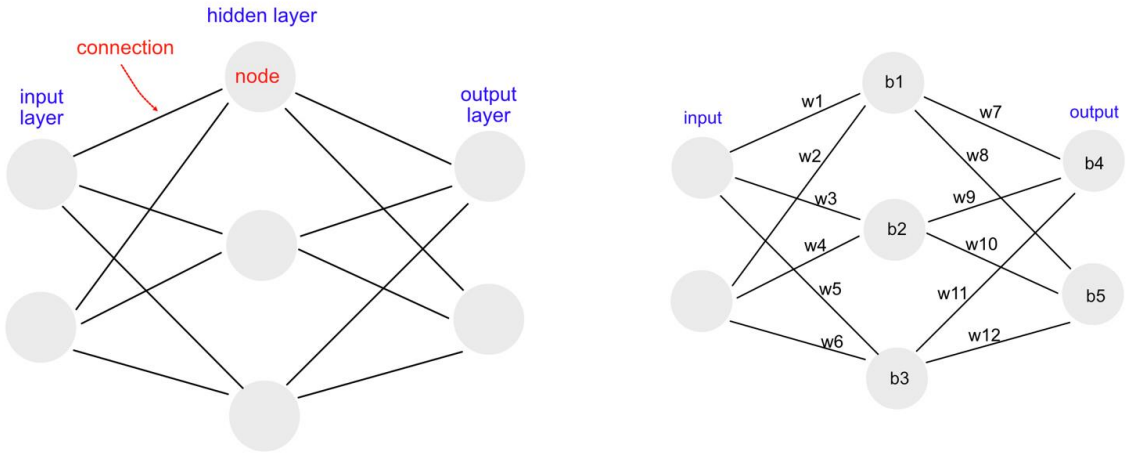
potential of LLM embeddings to act as a broad integrative platform for understanding visual cognition. It's especially interesting here then that LLM-trained ANNs, notably, often outperform alternative neuro-AI models in predicting brain activity, even when trained with fewer images.

A different comparison

Neural networks are inspired by the way the real brain function (lower left figure). They are simplified models of how *neurons* in our skulls (and spines) connect to each other. Whether it's the very recently published mouse connectome[5] or BRAIN data[6] (think of it as a cerebral census) with new results released last week[7], we are gradually piecing all these connections together. In an artificial neural network, the neurons are called nodes and they are arranged in layers. Each node is connected to nodes in the layer next to it. A neural network works by taking an input and converting it into an output. While this may seem basic, the output could represent something very significant. For example, the input could be the picture of a face and the output could be the name of the person in the picture. Training is an integral part of how neural networks work. During training, the network is fed inputs for which the correct outputs are known. For example, pictures of faces and names. The system compares the actual output to the correct answer, then goes back and weights connections within the network to amplify what is correct and diminish what is wrong. The process is then repeated many, many times. The training is complete when the network takes inputs and turns them into the correct outputs most of the time. The system is then ready to be unleashed on real-world data. For example, we feed 100 images of 10 people into the network, asking for their names. Every time the system spits out the wrong name, the internal adjustments become incrementally more accurate. Eventually the network's multitude of connections reflect the differences in the faces of 10 people based on the training images. The beauty of a trained neural network is that it contains generalised knowledge, not yes/no rules: if the system is fed a new picture of one of the 10 people, it will be able to identify them. Modern neural networks (lower right figure) are often sized based on how many "parameters" they have. A parameter is a number: either the "weight" of a connection or the "bias" on a node. In this example, there are 12 weights and 5 biases, for a total of 17 parameters.

[5] Ledford, Heidi. "AI Researchers Are Learning How to Share." Nature, February 14, 2025. https://www.nature.com/articles/d41586-025-00908-4.
[6] Nature Machine Intelligence Collection: Language Models and the Brain. Nature. Accessed November 25, 2025. https://www.nature.com/collections/gjdefhadcj.
[7] Naddaf, Miryam. "First-Ever Atlas of Brain Development Shows How Stem Cells Turn into Neurons." Nature, November 6, 2025. https://www.nature.com/articles/d41586-025-03641-0.

Figure 1: Structure of a Neural Network



The latest generation of large language models, such as GPT-5, have further increased their computational scale, reportedly reaching between 1.7 to 1.8 trillion parameters for dense model variants and tens of trillions of total parameters in mixture-of-experts configurations, or thereabouts. These models span more than 120 layers, and it's reasonable to estimate that GPT-5's dense network contains at least 11–15 million individual computational nodes. This is a remarkable engineering accomplishment, but when compared to the human brain—with its roughly 86 billion neurons—the biological system is still thousands of times larger. At the level of connection points, the brain's 100 trillion synapses dwarfs GPT-5's parameter count: even in its largest configurations, modern AI is only approaching a few percent of the network complexity found in a human cortex.

Yet sheer parameter counts can be misleading. The brain is a highly adaptive system—continually rewiring, involving interactive chemical signalling, and supported by vast biological infrastructure. Neurotransmitters such as dopamine and helper cells like astrocytes and microglia multiply the system's complexity far beyond what even the most advanced machine-learning architectures can approximate. If one hypothetically built an LLM with 86 billion nodes (roughly matching the brain's neuron count), the number of weights required—perhaps 60 quadrillion parameters—would demand computing resources on a scale far exceeding national energy grids. Even training GPT-5 required tens of thousands of megawatt hours of electricity, while a human brain operates at merely 20 watts. Over an 18-year developmental period, our brains consume vastly less energy than any comparable AI system. While I could drive into the power brain vs LLM debate in more detail another time, there are some truly fascinating papers on this topic. Clearly scale just matters: the bigger the LLM the more accurate the neuronal representation[8].

Despite these dramatic differences, GPT-5 and its predecessors demonstrate how much can be achieved with architectures that remain primitive relative to biological brains. A modern LLM still utilizes fewer nodes than a frog's nervous system but has shown extraordinary fluency in language generation, reasoning and multimodal learning. Innovations like hybrid model routing, modular neural plugins, and sparse attention have allowed GPT-5 to achieve more with less, reflecting ongoing advances in efficiency and architectural flexibility. This progress illustrates that parameter scaling alone does not capture intelligence: future models will likely integrate learning into their

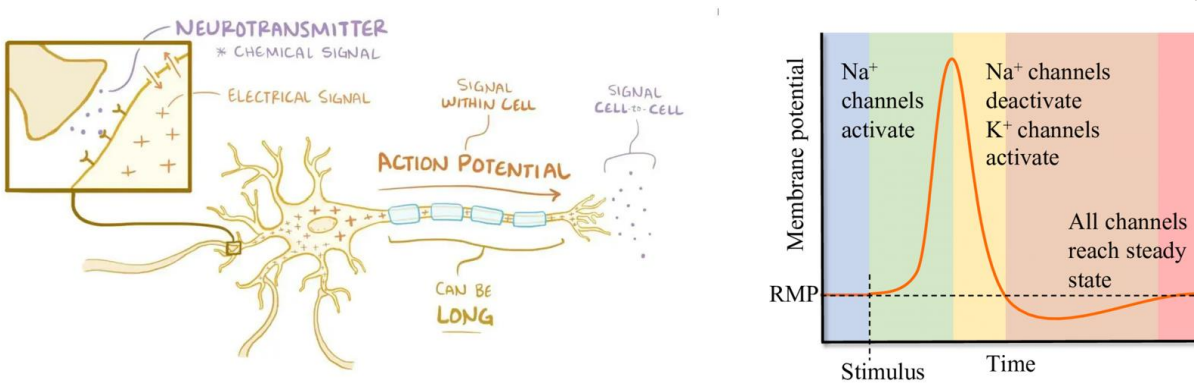
[8] Tang, Haiguang, Yiyuan Zhang, Yiyang Li, and Yunzhe Liu. "Brain-Inspired Multimodal Contrastive Learning of Representations for Vision and Language." Proceedings of the National Academy of Sciences of the United States of America (PNAS) 121, no. 1 (January 2, 2024): e2314298120. https://pmc.ncbi.nlm.nih.gov/articles/PMC11244877/.

operational phase, further blurring the distinction between training and inference, much like real brains. Fundamentally, human perception is irreducibly embodied and interpretive, integrating lived experience and conceptual understanding, as argued by philosophers such as Wittgenstein and Merleau-Ponty. AI models like GPT-5 build knowledge through statistical abstraction and pattern matching—not genuine comprehension—remixing vast training data without true embodiment. Current AI achievements remain extraordinary but also highlight the limitations inherent in disembodied learning and the gap yet to be bridged between artificial and biological cognition.

Chemistry vs electricity

Neurons act as biological transducers, converting chemical signals into electrical impulses and back in an intricate cycle central to all brain function. When a neuron receives a chemical message in the form of neurotransmitter molecules binding to receptors on its membrane, this triggers the opening of ion channels, allowing specific ions (notably sodium ions) to flow across the membrane. The resulting shift in electrical charge—the action potential—propagates rapidly along the axon to the neuron's end, where it once again triggers the release of neurotransmitters, passing the signal to the next neuron. This remarkable interplay takes advantage of the molecular machinery in the neuron's membrane, driven by finely tuned gradients of charged particles and an energy-consuming sodium-potassium pump that resets the system after each firing:

Figure 2: Neuronal Signalling and Membrane Potential Changes

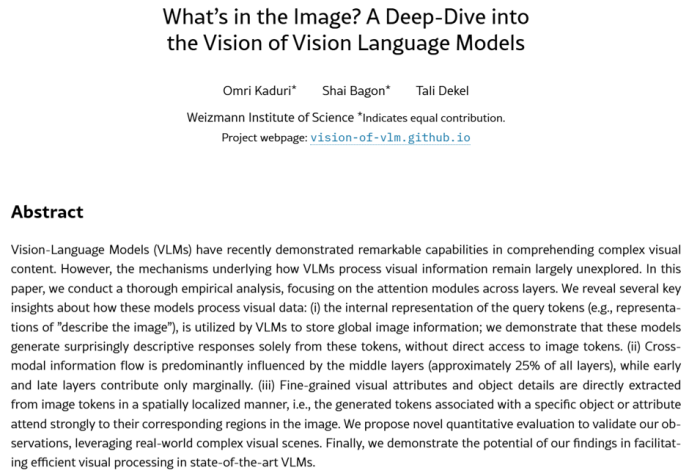


A long-standing debate exists over whether the "electricity" in the brain is digital or analogue. The answer is nuanced: the action potential itself is an all-or-nothing event, a spike that closely resembles digital signalling in being discrete in amplitude, but the frequency and timing of these spikes can be modulated in a continuous (analogue) fashion. Between actual spikes (digital), neurons also integrate inputs as graded potentials, continuous voltage shifts that sum up excitatory and inhibitory influences before the neuron decides to fire. Furthermore, many aspects of neurotransmitter release, synaptic strength, and even axonal propagation preserve continuous, analogue information alongside spiking events, leading researchers to conclude that the brain uses a hybrid model—combining both analogue and digital codes.

Taking the comparison further, digital computers—including all current LLMs—operate solely in the digital realm: their neurons (nodes) transmit information as numbers represented in discrete steps, whether binary or floating-point values. This equips them for speed, accuracy, and reproducibility, but requires tremendous energy and can lack the graceful, continuous adaptability intrinsic to biological systems. Analogue AI, by contrast, is a growing research field seeking to mimic these features, utilizing neuromorphic circuits and memristors to replicate spiking and the graded

dynamics of real neurons—but even these efforts are only approximations of what the brain achieves naturally, with efficiency and plasticity.

Going back to the vision discussion, for LLMs images are initially treated as arrays of pixel values. These models—often using architectures like Vision Transformers (ViT) or CLIP—divide the image into regular patches (e.g., 16×16 pixels). Each patch's pixel data is flattened and projected into an embedding vector, yielding a grid of tokens that is passed through deep transformer layers. As the image propagates through the model, these tokens are transformed and contextualized within high-dimensional space, gradually acquiring abstract features such as shapes or patterns that relate to object and scene categories, but starting from undifferentiated pixel-level input. In contrast, pioneering work by Hubel and Wiesel demonstrated that visual processing in the early brain is fundamentally biological and hierarchical: neurons in the retina and primary visual cortex respond selectively to simple features such as edges and orientations. As these signals propagate through successive layers, more complex shapes and objects are recognized. The coding is not an undifferentiated "mass of pixels," but emerges through highly specialized, feature-sensitive circuits that extract salient edges, contours, and gradually more abstract visual elements, providing an efficient and biologically optimized decomposition of scenes. Thus, while LLMs begin by slicing a picture into pixel patches as per the paper lower left and use learned statistical weights to reconstruct meaning, the brain exploits evolved, hardwired circuitry that directly encodes essential properties like directionality, contrast, and motion long before higher-level semantic interpretation. Both systems ultimately abstract from basic inputs to reach interpretive, conceptual representations, but the biological pathway has a deeply structured trajectory from the outset—one that inspired modern approaches to computer vision yet remains distinctly more efficient and context-aware than current artificial systems. The previously discussed paper provides a vision-LLM picture comparison:



What's in the Image? A Deep-Dive into the Vision of Vision Language Models

Omri Kaduri* Shai Bagon* Tali Dekel

Weizmann Institute of Science *Indicates equal contribution.

Project webpage: vision-of-vlm.github.io

Abstract

Vision-Language Models (VLMs) have recently demonstrated remarkable capabilities in comprehending complex visual content. However, the mechanisms underlying how VLMs process visual information remain largely unexplored. In this paper, we conduct a thorough empirical analysis, focusing on the attention modules across layers. We reveal several key insights about how these models process visual data: (i) the internal representation of the query tokens (e.g., representations of "describe the image"), is utilized by VLMs to store global image information; we demonstrate that these models generate surprisingly descriptive responses solely from these tokens, without direct access to image tokens. (ii) Cross-modal information flow is predominantly influenced by the middle layers (approximately 25% of all layers), while early and late layers contribute only marginally. (iii) Fine-grained visual attributes and object details are directly extracted from image tokens in a spatially localized manner, i.e., the generated tokens associated with a specific object or attribute attend strongly to their corresponding regions in the image. We propose novel quantitative evaluation to validate our observations, leveraging real-world complex visual scenes. Finally, we demonstrate the potential of our findings in facilitating efficient visual processing in state-of-the-art VLMs.

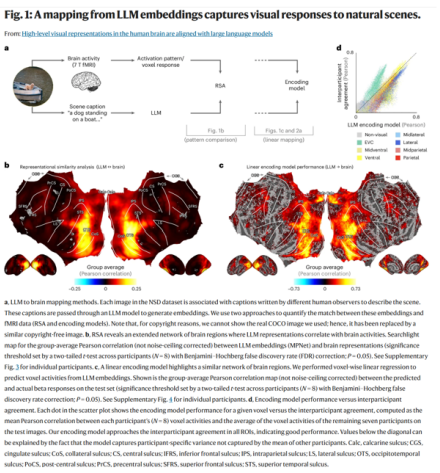


Fig. 1: A mapping from LLM embeddings captures visual responses to natural scenes.

From: High level visual representations in the human brain are aligned with large language models

Source: Kaduri, Omri, Shai Bagon, and Tali Dekel. "What's in the Image? A Deep-Dive into the Vision of Vision Language Models." arXiv preprint arXiv:2411.17491v1, November 26, 2024. https://arxiv.org/html/2411.17491v1.

Doerig, Adrien, Tim C. Kietzmann, and Nikolaus Kriegeskorte. "High-Level Visual Representations in the Human Brain Are Aligned with Large Language Models." Nature Machine Intelligence, August 7, 2025. Figure 1. https://www.nature.com/articles/s42256-025-01072-0/figures/1

From the standpoint of raw efficiency, adaptability, and multiplexing advantages, the brain's hybrid analogue-digital approach is superior, allowing rich sensory input, memory formation, and continuous adjustment over a power budget smaller than a household lightbulb. However, digital systems are peerless in rapid, exact, large-scale numerical calculation and scalability—attributes that have enabled astonishing progress in artificial intelligence. It's obviously tempting and rather obvious to state that the leading edge likely lies in integration: learning from the strengths of

brain-like analogue-digital hybrid signalling to build future AI systems that are smarter, more efficient, and more adaptable.

Conclusion

The contest between advanced LLMs and the human brain brings into relief the philosophical questions posed by Hume and Kant. Hume, emphasizing the mind as a blank slate shaped by the constant conjunction of empirical impressions, would see in modern AI a system that skilfully assembles patterns from its vast troves of sensory and linguistic data, building complex associations but lacking an intrinsic framework for causality or meaning. In both brains and machines, intelligence arises through experience, yet for Hume, even the most advanced LLMs remain fundamentally tied to correlation—reflecting statistical regularities, but not genuine understanding, agency, or intuition.

Kant, in contrast, insisted that human cognition is not merely receptive but actively structured by innate categories such as space, time, and causality—scaffolding sensory data into coherent, meaningful experience. The brain's continuous negotiation of chemistry and electricity, its analogue-digital hybrid signalling, and its capacity for meaning-making and judgment reflects a deeply Kantian organization: data are not raw, but always arranged by faculties that transcend empirical input. Thus, while LLMs like GPT-5 exemplify the Humean ideal of flexible, experience-driven systems, they lack the transcendental structures that anchor human comprehension, autonomy, and creativity. Ultimately, neither alone suffices; the future of AI may depend on synthesizing the adaptive strengths of Humean learning with the concept-generating, meaning-infusing capacities celebrated by Kant.