

How LLMs Can Read Your Mind

Imagine if AI could tell us what you were looking at, purely from a brain scan - or even tell us what our dog was thinking. This might sound like science fiction, but it might not be as far off as you think.

Researchers set out to answer a striking question: could an AI large language model (LLM) that has never seen a single image explain the way the human brain represents what we see?

In a study by Doerig et al., the authors found that an existing LLM trained purely on huge amounts of text with no image data involved whatsoever could represent visual scenes based only on the captions describing them in a strikingly similar way to how the human brain represents those same scenes when actually looking at them. Through this text-only training the model had picked up rich statistical relationships about how objects, actions, and contexts fit together in the world - which is exactly the kind of contextual, relational information the human brain itself extracts when making sense of a visual scene. This may explain why the two end up representing things so similarly.

To test this idea, they took human-written captions describing natural scenes (things like "a dog sitting on a couch") and fed them through the LLM which converts text into a kind of mathematical fingerprint (what researchers call an "embedding"). Separately, they scanned the brains of 8 participants using fMRI while those participants viewed the actual photos the captions described. Interestingly, they found that these embeddings could predict which regions of the brain would activate when a participant looked at a specific photo - essentially reverse-engineering the brain's visual response just from a caption.

The LLM's embedding for a caption closely matched the brain's activity pattern when a participant looked at the matching photo - even though the model had no visual experience whatsoever and had never encountered these images or this brain data. The match was strong enough that, working backwards, the researchers could take a participant's brain scan alone, with no image data involved at any stage, and generate an accurate caption for what they'd been looking at - correctly producing something like "a dog sitting on a couch".

How LLMs Match & Decode Brain Activity

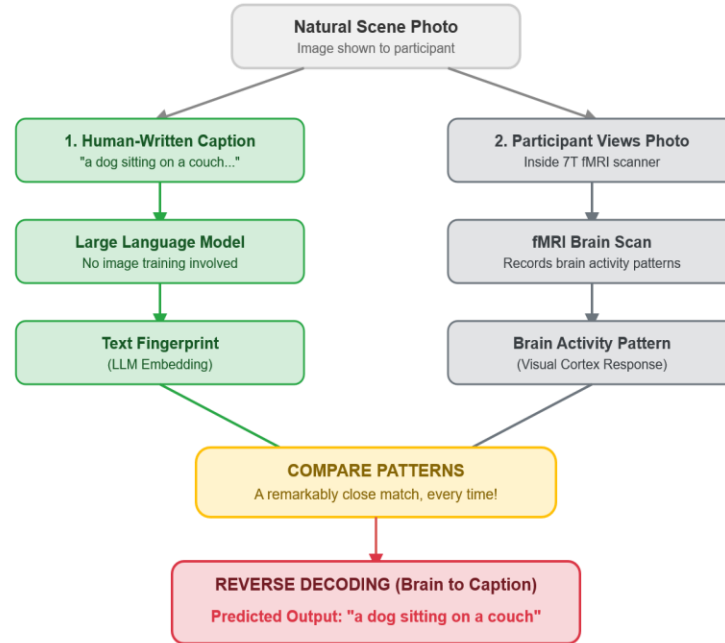


Figure 1: How the study worked, from photo to caption and brain scan, and back again.

Even more striking: when the researchers trained a new AI model to predict these LLM embeddings directly from raw images, it outperformed a wide range of leading vision models, including some trained on hundreds of millions of images - despite being trained on only about 48,000. Most of those other models had simply been trained to recognize what object is in a photo (a dog, a couch, etc.). Whereas this new model was trained to predict the LLM's richer caption embedding - one that already captures the objects, their relationships, and the context around them. That richer target seems to be what let it learn so much from so little data. This finding challenges the idea that bigger training datasets are always better for AI performance. Sometimes, giving a model the right target to aim for matters more than how much data you throw at it.

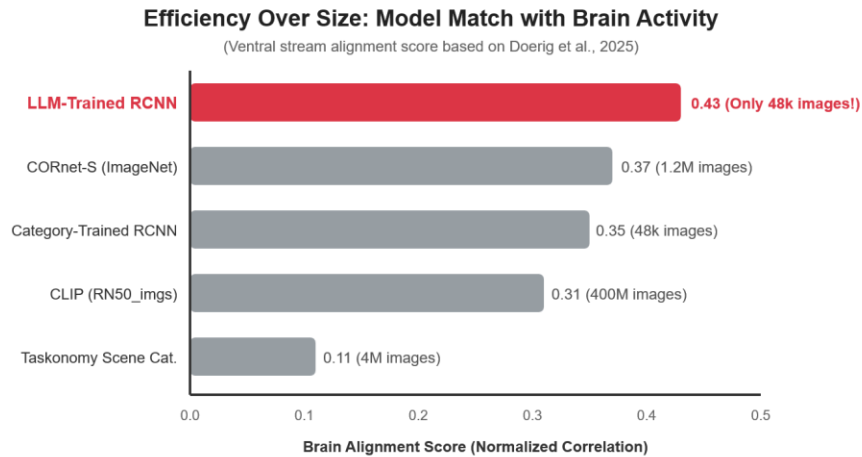


Figure 2: Illustrative recreation of Fig. 4e from Doerig et al. (2025), showing approximate model performance vs. training set size. Exact values are not published in the paper; this chart is simplified for clarity.

This research opens up exciting avenues for future work: does the way our brain represents visual information generalize to other senses, like hearing? And does it generalize to other species? If so, this approach could help us better understand how animals interpret the world around them - or even, one day, help us understand what our dogs are really thinking.

Source: Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2025). High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7, 1220–1234. <https://doi.org/10.1038/s42256-025-01072-0>