Reading machines

https://doi.org/10.1038/s43588-025-00905-7

Larger language models better align with the reading brain

Samuel A. Nastase

Check for updates

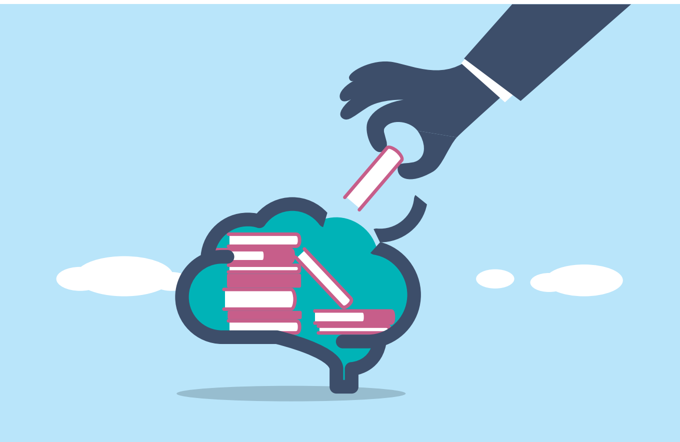
A systematic comparison of large language models suggests that larger models align better with both human behavior and brain activity during natural reading. Instruction tuning, however, does not yield a similar benefit.

Your brain effortlessly scans through the words on this page to construct the unique meaning of the text I've written. Current words affect the meaning of future words; the current sentence influences the meaning of future sentences. How does the brain manage to piece all of this together? Changjiang Gao, Zhengwu Ma and colleagues[1], writing in *Nature Computational Science*, look for answers in artificial neural network models for natural language. They leverage recent advances in language modeling to determine what features of language models best align with human reading.

Large language models (LLMs) are deep neural networks that have been trained to process real-world language and output useful, fluent responses. The foundation of modern LLMs is a remarkably simple learning objective: predicting the next word (or sub-word token) in massive corpora of online text. At the core of LLM architecture is the 'self-attention' circuit: for any given word, internal components of the model 'attend' to previous words in the text in order to better to capture the unique, context-specific meaning of the current word[2]. The model effectively learns to 'look back' at previous words to make sure its representation of the current word reflects the preceding context. Exactly what the model looks back at, and how the prior context sculpts the meaning of the current word, are things the model learns — all in pursuit of better next-word prediction. Recent work has shown that the internal representations of LLMs are more closely aligned with human brain activity than earlier classes of language model[3].

Language models have become remarkably more fluent, conversational and helpful over the past five years. Although there have been many engineering developments, there are two factors that appear to be particularly important. First, language models have become larger and larger over the years — with many more layers and more parameters. As language models get larger, they better capture the nuances of human language behavior[4]. Interestingly, larger language models also appear to align better with human brain activity[5,6]. Second, many modern language models are fine-tuned on the basis of human feedback, typically using examples of instructions paired with human-curated responses[7]. These instruction-tuned models yield outputs that are more useful and better preferred by human users. However, it has remained unclear, from a cognitive perspective, whether these instruction-tuned models align better with human behavior and brain activity than those trained on next-word prediction alone[8,9].

With this in mind, Gao and colleagues set out to test how these two factors affect model–human alignment during naturalistic reading.



CREDIT: WENMEI ZHOU / DIGITALVISION VECTORS / GETTY IMAGES

To do so, they used concurrent eye-tracking and functional magnetic resonance imaging (fMRI) to measure gaze and brain activity simultaneously while human subjects (50 native English-speakers) read scientific texts sentence by sentence. The authors then supplied the same texts sentence by sentence to several different LLMs and extracted the attention patterns from each model. Critically, they systematically tested models of increasing size, with and without instruction tuning, in terms of their alignment to human behavior and brain activity.

A large portion of prior work using LLMs to model human brain activity has used spoken narrative stimuli, in which subjects simply sit back and listen to stories. Reading, on the other hand, is a much more active process than listening. Readers move their eyes from word to word at their own unique pace. From an experimental perspective, this makes reading more difficult to study than listening. When reading, humans tend to rapidly glance back at previous words — these regressive saccades go against the typical direction of reading (left to right in English) and account for 15–25% of saccades[10]. Although the role of these saccades in language comprehension is still not fully understood, they may serve as a behavioral signature of the brain's efforts to resolve the present meaning of a text based on previous words. (Does this remind you of something you have already read in this manuscript? Did your eyes flit back to check?) The authors took advantage of this parallel between human reading behavior and the internal self-attention mechanism of LLMs to quantify model–human alignment.

First, Gao, Ma and colleagues tested how well a model's internal attention patterns — that is, which words the model tends to 'look back' at — match regressive saccades in human reading behavior. They found that larger models predicted human reading behavior better, but instruction-tuned models did not outperform their counterparts (same size but without instruction tuning). Second, the authors tested how well the model's internal attention patterns match fluctuations in brain activity corresponding to the same regressive saccades. Again, they found that larger models are better aligned with brain activity, but that instruction tuning did not improve model–brain alignment. Note that these scaling effects in model–human alignment are not simply due to overfitting in the alignment process: the models are evaluated on left-out stimuli to mitigate overfitting and the same effects are observed even when larger models are reduced to matching dimensionality[5,6]. Larger models appear to learn structures of language that smaller models cannot.

Overall, these findings confirm that larger models align better with human reading — but indicate that instruction tuning, despite yielding more useful models, does not appear to bring them closer to human cognition. Further work is needed to determine whether other kinds of fine-tuning may improve model–human alignment, or whether instruction tuning may enhance alignment for particular kinds of natural language tasks.

Why would simply making models larger bring them closer to humans? Scientists usually prefer simpler, more easily interpretable models and explanations. Taking a historical example from astronomy, LLMs might be accused of adding epicycles, in which a more parsimonious explanation of language processing remains hidden. Linguists have developed very elegant rules for describing many regularities of language. The patterns of real-world language, however, are incredibly rich, with both rule-like regularities and all manner of irregularities and contextual inflections. Despite the appeal of symbolic, rule-based models of linguistic structure, they have not scaled up to holistic, full-fledged language processing. Such models do not explain how all the remarkably diverse structures of language can be unified in a 'language' of neural activity, or how these structures can be obtained (whether through evolution or learning).

LLMs, on the other hand, do not appeal to any of the constructs of formal linguistics — despite being able to reproduce essentially all of the structures and patterns of natural language in generating fluent, meaningful responses. Instead, they encode all of these structures in a continuous, high-dimensional embedding space. They rely on neural population codes and a simple statistical learning algorithm. It's the scale of these models that makes them expressive enough to accommodate the rich contextual structure of everyday language. Despite the complexity of what they learn, these models are deceptively simple: for example, the self-attention circuit is famously summarized in a one-line equation. There is an unusual elegance in building a learning machine from which so many of the intricacies of language emerge simply by running it up against the structure of everyday language. This may give the linguist pause, but it's an exciting moment for those of us pursuing a computational neuroscience of natural language.

Samuel A. Nastase [1,2] ✉
[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. [2]Department of Psychology and Center for Computational Language Sciences, University of Southern California, Los Angeles, CA, USA.
✉e-mail: sam.nastase@gmail.com

Published online: 12 November 2025

References

1. Gao, C. et al. *Nat. Comput. Sci.* https://doi.org/10.1038/s43588-025-00863-0 (2025).
2. Vaswani, A. et al. *Adv. Neural Inf. Process. Syst.* **30**, 6000–6010 (2017).
3. Kumar, S. et al. *Nat. Commun.* **15**, 5523 (2024).
4. Kaplan, J. et al. Preprint at https://arxiv.org/abs/2001.08361 (2020).
5. Antonello, R., Vaidya, A. & Huth, A. *Adv. Neural Inf. Process. Syst.* **36**, 21895–21907 (2023).
6. Hong, Z. et al. *eLife* **13**, RP101204 (2024).
7. Ouyang, L. et al. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
8. Kuribayashi, T., Oseki, Y. & Baldwin, T. Psychometric predictive power of large language models. In *Findings of the Association for Computational Linguistics (NAACL 2024)* (eds Duh, K. et al.) 1983–2005 (ACL, 2024).
9. Aw, K. L., Montariol, S., AlKhamissi, B., Schrimpf, M. & Bosselut, A. Instruction-tuning aligns LLMs to the human brain. In *Conference on Language Modeling (COLM 2024)* https://openreview.net/forum?id=nXNN0x4wbl (2024).
10. Rayner, K. *Psychol. Bull.* **124**, 372–422 (1998).

Competing interests

The author declares no competing interests.