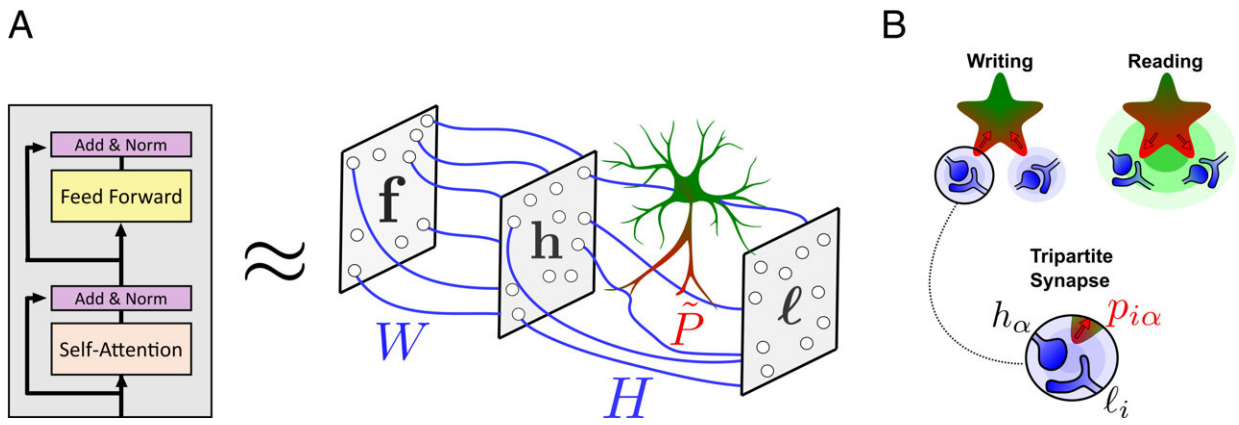
# AI models are powerful, but are they biologically plausible?

August 15 2023, by Adam Zewe




(A) A high-level overview of the proposed neuron–astrocyte network. The Transformer block is approximated by a feed-forward network with an astrocyte unit that ensheaths the synapses between the hidden and last layers (matrix H). Data are constantly streamed into the network. (B) During the writing phase the neuron-to-neuron weights are updated using Hebbian learning rule and the neuron-to-astrocyte weights are updated using a presynaptic plasticity rule. During the reading phase, the data are forwarded through the network, and the astrocyte modulates the synaptic weights H. Credit: *Proceedings of the National Academy of Sciences* (2023). DOI: 10.1073/pnas.2219150120

Artificial neural networks, ubiquitous machine-learning models that can be trained to complete many tasks, are so called because their architecture is inspired by the way biological neurons process information in the human brain.

About six years ago, scientists discovered a new type of more powerful neural network model known as a transformer. These models can achieve unprecedented performance, such as by generating text from prompts with near-human-like accuracy. A transformer underlies AI systems such as ChatGPT and Bard, for example. While incredibly effective, transformers are also mysterious: Unlike with other brain -inspired neural network models, it hasn't been clear how to build them using biological components.

Now, researchers from MIT, the MIT-IBM Watson AI Lab, and Harvard Medical School have produced a hypothesis that may explain how a transformer could be built using biological elements in the brain. They suggest that a biological network composed of neurons and other brain cells called astrocytes could perform the same core computation as a transformer.

Recent research has shown that astrocytes, non-neuronal cells that are abundant in the brain, communicate with neurons and play a role in some physiological processes, like regulating blood flow. But scientists still lack a clear understanding of what these cells do computationally.

With the new study, published this week in *Proceedings of the National Academy of Sciences*, the researchers explored the role astrocytes play in the brain from a computational perspective, and crafted a mathematical model that shows how they could be used, along with neurons, to build a biologically plausible transformer.

Their hypothesis provides insights that could spark future neuroscience research into how the human brain works. At the same time, it could help machine-learning researchers explain why transformers are so successful across a diverse set of complex tasks.

"The brain is far superior to even the best artificial neural networks that

we have developed, but we don't really know exactly how the brain works. There is scientific value in thinking about connections between biological hardware and large-scale artificial intelligence networks. This is neuroscience for AI and AI for neuroscience," says Dmitry Krotov, a research staff member at the MIT-IBM Watson AI Lab and senior author of the research paper.

Joining Krotov on the paper are lead author Leo Kozachkov, a postdoc in the MIT Department of Brain and Cognitive Sciences; and Ksenia V. Kastanenka, an assistant professor of neurobiology at Harvard Medical School and an assistant investigator at the Massachusetts General Research Institute.

## A biological impossibility becomes plausible

Transformers operate differently than other neural network models. For instance, a recurrent neural network trained for natural language processing would compare each word in a sentence to an internal state determined by the previous words. A transformer, on the other hand, compares all the words in the sentence at once to generate a prediction, a process called self-attention.

For self-attention to work, the transformer must keep all the words ready in some form of memory, Krotov explains, but this didn't seem biologically possible due to the way neurons communicate.

However, a few years ago scientists studying a slightly different type of machine-learning model (known as a Dense Associated Memory) realized that this self-attention mechanism could occur in the brain, but only if there were communication among at least three neurons.

"The number three really popped out to me because it is known in neuroscience that these cells called astrocytes, which are not neurons,

form three-way connections with neurons, what are called tripartite synapses," Kozachkov says.

When two neurons communicate, a presynaptic neuron sends chemicals called neurotransmitters across the synapse that connects it to a postsynaptic neuron. Sometimes, an astrocyte is also connected—it wraps a long, thin tentacle around the synapse, creating a tripartite (three-part) synapse. One astrocyte may form millions of tripartite synapses.

The astrocyte collects some neurotransmitters that flow through the synaptic junction. At some point, the astrocyte can signal back to the neurons. Because astrocytes operate on a much longer time scale than neurons—they create signals by slowly elevating their calcium response and then decreasing it—these cells can hold and integrate information communicated to them from neurons. In this way, astrocytes can form a type of memory buffer, Krotov says.

"If you think about it from that perspective, then astrocytes are extremely natural for precisely the computation we need to perform the attention operation inside transformers," he adds.

## Building a neuron-astrocyte network

With this insight, the researchers formed their hypothesis that astrocytes could play a role in how transformers compute. Then they set out to build a mathematical model of a neuron-astrocyte network that would operate like a transformer.

They took the core mathematics that comprise a transformer and developed simple biophysical models of what astrocytes and neurons do when they communicate in the brain, based on a deep dive into the literature and guidance from neuroscientist collaborators.

Then they combined the models in certain ways until they arrived at an equation of a neuron-astrocyte network that describes a transformer's self-attention.

"Sometimes, we found that certain things we wanted to be true couldn't be plausibly implemented. So, we had to think of workarounds. There are some things in the paper that are very careful approximations of the transformer architecture to be able to match it in a biologically plausible way," Kozachkov says.

Through their analysis, the researchers showed that their biophysical neuron-astrocyte network theoretically matches a transformer. In addition, they conducted numerical simulations by feeding images and paragraphs of text to transformer models and comparing the responses to those of their simulated neuron-astrocyte network. Both responded to the prompts in similar ways, confirming their theoretical model.

The next step for the researchers is to make the leap from theory to practice. They hope to compare the model's predictions to those that have been observed in biological experiments, and use this knowledge to refine—or possibly disprove—their hypothesis.

In addition, one implication of their study is that astrocytes may be involved in long-term memory, since the network needs to store information to be able act on it in the future. Additional research could investigate this idea further, Krotov says.

"For a lot of reasons, astrocytes are extremely important for cognition and behavior, and they operate in fundamentally different ways from neurons. My biggest hope for this paper is that it catalyzes a bunch of research in computational neuroscience toward glial cells, and in particular, astrocytes," adds Kozachkov.

**More information:** Leo Kozachkov et al, Building transformers from neurons and astrocytes, *Proceedings of the National Academy of Sciences* (2023). DOI: 10.1073/pnas.2219150120

*This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.*

Provided by Massachusetts Institute of Technology