PNAS RESEARCH ARTICLE | NEUROSCIENCE COMPUTER SCIENCES OPEN ACCESS

Building transformers from neurons and astrocytes

Leo Kozachkov[a,b,1], Ksenia V. Kastanenka[c], and Dmitry Krotov[a,1]

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received November 9, 2022; accepted June 22, 2023

Glial cells account for between 50% and 90% of all human brain cells, and serve a variety of important developmental, structural, and metabolic functions. Recent experimental efforts suggest that astrocytes, a type of glial cell, are also directly involved in core cognitive processes such as learning and memory. While it is well established that astrocytes and neurons are connected to one another in feedback loops across many timescales and spatial scales, there is a gap in understanding the computational role of neuron–astrocyte interactions. To help bridge this gap, we draw on recent advances in AI and astrocyte imaging technology. In particular, we show that neuron–astrocyte networks can naturally perform the core computation of a Transformer, a particularly successful type of AI architecture. In doing so, we provide a concrete, normative, and experimentally testable account of neuron–astrocyte communication. Because Transformers are so successful across a wide variety of task domains, such as language, vision, and audition, our analysis may help explain the ubiquity, flexibility, and power of the brain's neuron–astrocyte networks.

neuroscience | astrocytes | Transformers | glia | artificial intelligence

Significance

Transformers have become the default choice of neural architecture for many machine learning applications. Their success across multiple domains such as language, vision, and speech raises the question: How can one build Transformers using biological computational units? At the same time, in the glial community, there is gradually accumulating evidence that astrocytes, formerly believed to be passive house-keeping cells in the brain, in fact play an important role in the brain's information processing and computation. In this work we hypothesize that neuron–astrocyte networks can naturally implement the core computation performed by the Transformer block in AI. The omnipresence of astrocytes in almost any brain area may explain the success of Transformers across a diverse set of information domains and computational tasks.

Astrocytes, one kind of glia, are a ubiquitous cell type in the central nervous system. It is empirically well established that astrocytes and neurons communicate with one another via feedback loops that span many spatial and temporal scales (1–3). These communications underlie a variety of important physiological processes, such as regulating blood flow to neurons (4) and eliminating debris (5). A rapidly growing body of evidence suggests that astrocytes also play an active and flexible role in behavior (6–12). However, a firm computational interpretation of neuron–astrocyte communication is missing.

Transformers, a particular type of artificial intelligence (AI) architecture, have become influential in machine learning (13) and, increasingly, in computational neuroscience (14–20). They are currently the choice model for tasks across many disparate domains, including natural language processing, vision, and speech (21). Interestingly, several recent reports suggested architectural similarities between Transformers and the hippocampus (15, 19) and cerebellum (18), as well as representational similarities with human brain recordings (14, 16, 20). However, unlike more traditional neural networks, such as convolutional networks (22) or Hopfield networks (23), which have a long tradition of biological implementations, Transformers are only at the beginning of their interpretation in terms of known biological processes.

We hypothesize that biological neuron–astrocyte networks can perform the core computations of a Transformer. In support of this hypothesis, we explicitly construct an artificial neuron–astrocyte network whose internal mechanics and outputs approximate those of a Transformer with high probability. The main computational element of our network is the tripartite synapse, the ubiquitous three-factor connection between an astrocyte, a presynaptic neuron, and a postsynaptic neuron (24). We argue that tripartite synapses can perform the role of normalization in the Transformer's self-attention operation. As such, neuron–astrocyte networks are natural candidates for the biological "hardware" that can be used for computing with Transformers.

The organization of this paper is as follows. We begin with two primers, which introduce the core concepts and notations: one on astrocyte biology and the other one on Transformers. Then, we describe our neuron–astrocyte network in detail and demonstrate the correspondence to Transformers through theory and simulations. We begin by establishing the correspondence for the models with shared weights and then show the general case. For completeness, we also derive a nonastrocytic mechanism for implementing Transformers biologically. Although, ultimately, it should be decided through experiments which of the two mechanisms is closer to biological reality, from the theoretical perspective we argue that astrocytes provide a more natural and parsimonious hypothesis for how Transformers might be implemented in the brain. We conclude with

Author affiliations: [a]Massachusetts Institute of Technology-International Business Machines, Watson Artificial Intelligence Laboratory, IBM Research, Cambridge, MA 02142; [b]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; and [c]Department of Neurology, MassGeneral Institute for Neurodegenerative Diseases, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02115

Author contributions: L.K. and D.K. designed research; L.K. and D.K. performed research; and L.K., K.V.K., and D.K. wrote the paper.

Competing interest statement: L.K. did his summer internship at IBM.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: leokoz8@mit.edu or krotov@ibm.com.

Published August 14, 2023.

a discussion on the intrinsic timescales of our biological Transformers, as well as potential future work.

Primer on Astrocyte Biology. Glial cells are the other major cell type in the brain besides neurons. The exact ratio of glia to neurons is disputed, but it is somewhere between 1:1 and 10:1 (25). The most well-studied type of glial cell is the astrocyte. A defining feature of astrocytes is that a single astrocyte cell forms connections with thousands to millions of nearby synapses (26). For example, a single human astrocyte can cover between 270,000 to 2 million synapses within a single domain (27). Astrocytes are mostly electrically silent, encoding information in the dynamics of intracellular calcium ions (Ca^{2+}). In most parts of the brain, neurons and astrocytes are closely intertwined. For example, in the hippocampus as many as 60% of all axon–dendrite synapses are wrapped by astrocyte cell membranes called processes (28). In the cerebellum, the number is even higher. This three-way arrangement (presynaptic axon, postsynaptic dendrite, astrocytes process) is so common that it has been given a name: the tripartite synapse (24).

Astrocyte processes contain receptors corresponding to the neurotransmitters released at the synaptic sites they ensheathe. For example, astrocytes in the basal ganglia are sensitive to dopamine, whereas in the cortex astrocytes are sensitive to glutamate (29). Despite being affected by the same presynaptic neurotransmitters, postsynaptic neurons and astrocytes respond very differently: Neurons primarily encode information using action potentials, but astrocytes encode information via elevations in free intercellular calcium. Importantly, neuron-to-astrocyte signaling can trigger a response in the opposite astrocyte-to-neuron direction thus establishing a feedback loop between neural cells and astrocytes. Astrocytes can either depress or facilitate synapses, depending on the situation (30). For example, astrocytes in the hypothalamus have been observed to multiplicatively scale the excitatory synapses they ensheathe by the same common factor (31).

Interestingly, there is also extensive astrocyte-to-astrocyte communication in the brain. Astrocytes form large-scale networks with one another (26). These networks are spatially tiled, with regular intercellular spacing of a few tens of micrometers (32). Unlike neurons, which communicate primarily with spikes, astrocytes communicate via calcium waves that propagate between their cell bodies, processes, and endfeet (33). These waves have speeds of a few tens of micrometers per second. It is thought that these waves could be used to synchronize neural populations and coordinate important neural processes (34).

Among this plethora of biological phenomena, the following four points will be important for our mathematical model:

- Most synapses in the brain are tripartite (presynaptic neuron, postsynaptic neuron, astrocyte process).
- There is a feedback loop between astrocyte processes and synapses. Astrocyte processes respond to presynaptic neural activity with an elevation in intracellular calcium ions (Ca^{2+}) and, in turn, release gliotransmitters which modulate synapses. This modulation can be either facilitating or depressing.
- The neuron $\rightarrow$ astrocyte signaling pathway is plastic.
- Nearby astrocyte processes can spatially average their Ca^{2+} levels.

Next, we introduce Transformers from the AI perspective, before proposing their biological implementation with astrocytes.

Primer on Transformers. Transformers (13) are a popular neural architecture used in many of the recent innovations in AI including Foundation Models (35), Generative Pre-trained Transformer-3 (GPT3) (36), Chat Generative Pre-trained Transformer (ChatGPT) (37), etc. Originally developed for natural language processing tasks, Transformers are taking over the leader boards in other domains too, including vision (38), speech, and audio processing (21). Initially, Transformers were developed as a means to overcome the shortcomings of recurrent neural networks (13). A major difference between these two architectures is as follows: while recurrent neural network process inputs one at a time, Transformers have direct access to all past inputs. Through their self-attention mechanism (described in detail shortly), Transformers can learn long-range dependencies between words in a sentence without having to recurrently maintain a hidden state over long time intervals. Among other computational benefits, this allows for more efficient parallelization during the training process and avoids the vanishing/exploding gradient problem (39–41). In the vision domain, Transformers have also achieved state-of-the-art results (38) surpassing convolutional neural networks. While the latter use hard-coded inductive biases enabling them to learn local correlations between pixels in the images plane, Transformers form long-range learnable dependencies in the image plane right away starting from the early layers of processing (42).

Although recurrent and convolutional neural networks admit straightforward biological interpretations, Transformers presently do not. The reason has to do with the Transformer's self-attention mechanism. In particular, the so-called self-attention matrix is computed by a) calculating all pairwise dot products between "tokens" (e.g., words in a sentence, patches in an image, etc), b) exponentiating these dot product terms, and then c) normalizing the rows of this matrix to sum to one. These operations are fundamentally nonlocal in time and space, which make them difficult to interpret in biological terms. Later on, we will show how astrocyte biology offers a biologically plausible solution to this dilemma.

Transformers are typically a composition of many Transformer "blocks." A typical Transformer block uses four basic operations: self-attention, feed-forward neural network, layer normalization, and skip connections. These operations are arranged in a certain way so that the entire block can learn relationships between the tokens, which represent the data. More formally, consider a sequence of N token embeddings. Each token can correspond to a word (or a part of the word) if the Transformer is used in the language domain or a patch of an image in the vision domain. Each embedding is of dimension d. The tokens are streamed into the network one by one (online setting), and the time of the token's presentation is denoted by t. The t^{th} embedding is given by a vector $\mathbf{x}_t \in \mathbb{R}^d$. Going forward, it will be helpful to collect these tokens into a single matrix, X:

$$X \equiv \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_N \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{d\times N}. \qquad [1]$$

In the Transformer block, each token is converted to a key, query, and value vector via a corresponding linear transformation: W_K, $W_Q \in \mathbb{R}^{D\times d}$ and $W_V \in \mathbb{R}^{d\times d}$. Here, D is the internal size of the attention operation. These transformations are optimized during training. The key, value, and query vectors are then collected into matrices, similarly to Eq. **1**:

$$\begin{aligned} \mathbf{k}_t &= W_K \mathbf{x}_t \\ \mathbf{v}_t &= W_V \mathbf{x}_t \\ \mathbf{q}_t &= W_Q \mathbf{x}_t \end{aligned} \quad \rightarrow \quad \begin{aligned} K &= W_K X \\ V &= W_V X, \\ Q &= W_Q X \end{aligned} \qquad [2]$$

After computing the key, value, and query matrices, the next major step in a Transformer is the self-attention operation, which allows the tokens to exchange information with each other. The self-attention matrix, SelfAttn(X), is an $N \times N$ matrix which contains information about all the pairwise interactions between tokens. At the core of the self-attention mechanism is the softmax function. Recall that the softmax function exponentiates the elements of a vector and then divides each element by the sum of these exponentials. Denoting column t of the self-attention matrix by attn(t), we have that

$$\text{attn}(t) = \sum_{i=1}^{N} \alpha_i(t)\mathbf{v}_i \quad \text{with} \quad \alpha_i(t) = \frac{e^{\mathbf{k}_i^T \mathbf{q}_t}}{\sum_{j=1}^{N} e^{\mathbf{k}_j^T \mathbf{q}_t}}.$$

Due to the softmax normalization, each column of the self-attention matrix can be interpreted as a convex combination of the value vectors. Given this definition as well as Eq. **2**, we can write the self-attention matrix compactly as:

$$\text{SelfAttn}(X) = V\text{softmax}(K^T Q), \qquad [3]$$

where here the softmax normalization is computed along the *columns* of $K^T Q$. The output of this self-attention operation is then passed along to a LayerNorm operation and a feed-forward neural network (FFN) that both act separately on each token (each column of its input), see Fig. 1. Recall that a LayerNorm scales each element of a vector by the mean and variance of all elements in the vector (43) and can be implemented in a biologically plausible manner (44). Without loss of generality, a single-headed attention Transformer is studied. In this case, the output of the full Transformer block may be written as a two-step process:

$$\begin{aligned} Y &= \text{LayerNorm}(\text{SelfAttn}(X) + X) \\ \text{Transformer}(X) &= \text{LayerNorm}(\text{FFN}(Y) + Y), \end{aligned} \qquad [4]$$

where FFN refers to a feedfoward network, applied to each token (i.e., each column of Y) separately and identically.

Biological Implementation of a Transformer Block

In order to gain theoretical insight into Transformers, it is common to tie the weights (45, 46). This tying can be within a single Transformer block, between blocks, or both. In this section, we will tie the weights within a single block but not between blocks. We will relax this weight sharing constraint in the later sections. In particular, we tie W_Q, W_K, W_V as follows:

$$W_Q = W_K = W, \quad W_V = I, \qquad [5]$$

for some arbitrary matrix W and the identity matrix, I. In general, we will not require that $d = D$. We include this constraint now to fully analyze the simplest version of our model that captures the essential elements of our argument. Without loss of generality, we will ignore layer normalization steps for now, returning to them in the section titled "General Case of Untied Weights."

Neuron–Astrocyte Network. A high-level overview of our circuit is shown in Fig. 1. The network consists of a perceptron with an input layer, a hidden layer, and an output layer (Fig. 1*A*). As in many associative memory systems, our network has distinct writing and reading operations (23, 47). In particular, our network alternates between writing and reading *phases* (Fig. 1*B*). The writing phase enables the circuit to store information about all the tokens; the reading phase enables any given token to interact with all the others. Recall that a difficulty with interpreting Transformers as biological circuits is that they require operations which are nonlocal in space and time. Having distinct writing and reading phases allows our network to resolve this temporal nonlocality. As we will see, the spatial nonlocality is resolved through the astrocyte unit.

The d-dimensional inputs are passed to the hidden layer with m units, as well as to the last layer via a skip connection (not shown in Fig. 1). The hidden layer applies a fixed nonlinearity to incoming inputs. The outputs of the hidden layer are passed to the last layer via a linear mapping $H \in \mathbb{R}^{d \times m}$. The synapses in the matrix H are triparite synapses, meaning that each of the md synapses is associated with an astrocyte process $p_{i\alpha}$. The Latin indices i, j are used to enumerate neurons in the first and last layers, while the Greek indices α, β are reserved for the hidden neurons. The strength of the synapse between a hidden neuron α and the output neuron i is denoted by $H_{i\alpha}$ and the activity of the astrocyte process that ensheaths this synapse is described by $p_{i\alpha}$. The layers are denoted from left to right as $\mathbf{f}$,$\mathbf{h}$,$\boldsymbol{\ell}$ (first, hidden,

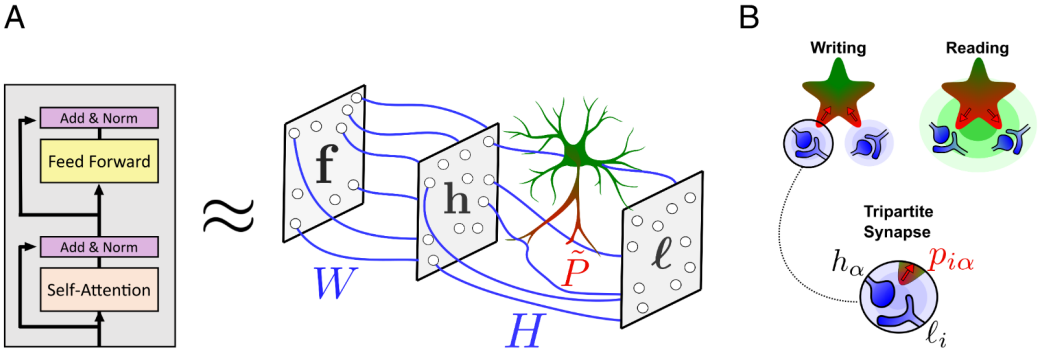

Fig. 1. (*A*) A high-level overview of the proposed neuron–astrocyte network. The Transformer block is approximated by a feed-forward network with an astrocyte unit that ensheaths the synapses between the hidden and last layers (matrix *H*). Data are constantly streamed into the network. (*B*) During 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*.

last), respectively. Our network is described by the following equations:

$$\begin{aligned} \mathbf{f} &= \mathbf{x} && \in \mathbb{R}^d \\ \mathbf{h} &= \phi(W\mathbf{f}) && \in \mathbb{R}^m \\ \boldsymbol{\ell} &= r(H \odot \tilde{P})\mathbf{h} + \mathbf{f} && \in \mathbb{R}^d, \end{aligned} \quad [6]$$

The scalar $r = \{0, 1\}$ stands for 'read' and is zero during the writing phase and unity during the reading phase. The symbol $\odot$ denotes the Hadamard product (element-wise multiplication) between two matrices. The matrix $\tilde{P} \in \mathbb{R}^{d \times m}$ captures the effect of the astrocyte processes and is defined as follows:

$$\tilde{P}_{i\alpha} = \frac{1}{p_{i\alpha}}$$

This inverse modulation of synaptic weights by astrocytes has been observed, for example, in studies involving tumour necrosis factor-alpha (TNF-α), wherein astrocytes will upscale synaptic weights in response to low neural activity and downscale weights in response to high neural activity. More generally, many studies have observed that astrocytes can both depress and facilitate synapses, depending on the situation (1, 48–51).

Neural Activation Function. The neural activation function ϕ plays a special role in our circuit. In order to match the exponential dot product in the Transformer's self-attention mechanism, we will require that ϕ be an approximate feature map for the exponential dot product kernel

$$\phi(\mathbf{x})^T \phi(\mathbf{y}) \approx e^{\mathbf{x}^T \mathbf{y}}, \quad [7]$$

There are many (indeed, infinitely many) activation functions which satisfy this condition. Several biologically plausible options come from the theory of random feature maps (52–54), and we will discuss them in detail later on. For now, we will simply assume that ϕ is chosen so that Eq. 7 is true. More generally, however, one can pick any ϕ such that $\phi(\mathbf{x})^T \phi(\mathbf{y}) \geq 0$ to yield a valid self-attention mechanism (55). Nevertheless, only particular choices of ϕ yield the softmax self-attention which is used in most Transformers at scale (13).

Astrocyte Process Dynamics. As discussed in the introduction, astrocyte processes are sensitive to presynaptic neural activity. To capture this mathematically, we assume that the astrocyte process Ca^{2+} response is linearly proportional to the presynaptic neuron activation h_α of neuron α in layer $\mathbf{h}$. The constant of proportionality between the astrocyte process activation and the presynaptic neural activity is denoted as $g_{i\alpha}$. This constant is in general different for every astrocyte process. Upon presentation of an embedded token to the network, astrocyte process $p_{i\alpha}$ initially responds with a local calcium elevation $g_{i\alpha} h_\alpha$. This Ca^{2+} response is then spatially averaged with the responses of other nearby astrocyte processes so that, after transients, the processes have the same value once a token is presented:

$$p_{i\alpha} = \frac{1}{md} \sum_{j=1}^{d} \sum_{\beta=1}^{m} g_{j\beta} h_\beta = p. \quad [8]$$

The neuron-to-astrocyte signaling pathway in our circuit is completely described by Eq. 8.

Writing Phase. During the writing phase, r is set to zero. Biologically, this condition could correspond to some global neuromodulator being released into the local environment, for example, acetylcholine, as suggested in refs. 17 and 56. Plugging $r = 0$, Eq. 6 becomes

$$\begin{aligned} \mathbf{f}_t &= \mathbf{x}_t \\ \mathbf{h}_t = \phi(W\mathbf{f}_t) &= \phi(\mathbf{k}_t) \\ \boldsymbol{\ell}_t &= \mathbf{f}_t = \mathbf{v}_t, \end{aligned} \quad [9]$$

where we have substituted in the definitions of the key, query, and value vectors given by Eq. 2, as well as the temporary weight-tying assumption given by Eq. 5. As the embedded tokens are passed into Eq. 9 sequentially, the weight matrix H is updated via Hebbian plasticity with a learning rate of $\frac{1}{m}$. Upon presentation of token t, the matrix H is

$$H_t = H_{t-1} + \frac{1}{m} \boldsymbol{\ell}_t \mathbf{h}_t^T \implies H = \frac{1}{m} V \phi^T(K),$$

where we have assumed that H is initially the zero matrix and substituted in the equalities in Eq. 9. At the same time that the neuron-to-neuron weights are updated via Hebbian plasticity, the neuron-to-astrocyte weights are updated via presynaptic plasticity. Upon presentation of token t, these weights are

$$g_{t,i\alpha} = g_{t-1,i\alpha} + \phi(W\mathbf{x}_t)_\alpha \implies g_{i\alpha} = \sum_{j=1}^{N} \phi(\mathbf{k}_j)_\alpha.$$

Note that as a consequence of the presynaptic plasticity, the weight $g_{i\alpha}$ does not depend on the index i. Therefore, we will only refer to the vector $\mathbf{g} \in \mathbb{R}^m$, which—through the presynaptic plasticity—is simply the sum over all token presentations of the hidden layer neural activations:

$$\mathbf{g} = \sum_{j=1}^{N} \phi(\mathbf{k}_j).$$

Reading Phase. During the reading phase, the read gate is set to $r = 1$ in Eq. 6, and the inputs are forwarded through the network. The astrocyte process activation value p, which according to Eq. 8 does not depend on indices i and α, is given by

$$p = \frac{d}{md} \mathbf{g}^T \mathbf{h} = \frac{1}{m} \sum_{j=1}^{N} \phi(\mathbf{k}_j)^T \phi(\mathbf{q}_t). \quad [10]$$

To obtain the last equality, we have used $\mathbf{h}_t = \phi(W\mathbf{x}_t) = \phi(\mathbf{q}_t)$. Plugging in all the steps of Eq. 6, we see that the last layer has the following output

$$\begin{aligned} \boldsymbol{\ell}_t &= \frac{1}{p} H \phi(\mathbf{q}_t) + \mathbf{x}_t = \frac{V \phi^T(K) \phi(\mathbf{q}_t)}{\phi(\mathbf{q}_t)^T \sum_{j=1}^{N} \phi(\mathbf{k}_j)} + \mathbf{x}_t \\ &\approx \sum_{i=1}^{N} \frac{e^{\mathbf{k}_i^T \mathbf{q}_t}}{\sum_{j=1}^{N} e^{\mathbf{k}_j^T \mathbf{q}_t}} \mathbf{v}_i + \mathbf{x}_t \\ &= \mathrm{attn}(t) + \mathbf{x}_t, \end{aligned} \quad [11]$$

where we have used the assumption that ϕ is an approximate feature map for the exponential dot product, given by Eq. 7. If we

compute ℓ_t for every token $\mathbf{x}_t$ and stack the results column-wise into a matrix L, we can conclude that the output of our neuron–astrocyte circuit is approximately the output of the Transformer's self-attention, plus the necessary residual connection:

$$L \approx \text{SelfAttn}(X) + X. \quad [12]$$

Random Feature Activations. As mentioned above, in order to approximate the softmax attention, we require that ϕ is a feature map for the exponential dot product. This is the idea behind linear Transformer architectures (55) such as Performers (53) and Random Feature Attention (54). We will now discuss two biologically plausible options for such a feature map. The first relies on a well-known result in kernel approximation theory (52), which is that the radial basis function (RBF) kernel can, with high probability, be approximated very well using random projections and cosines

$$\phi(\mathbf{x}) = \sqrt{\frac{2}{m}}\exp\Big(\frac{||\mathbf{x}||^2}{2}\Big)\cos(\Pi\mathbf{x} + \mathbf{b}), \quad [13]$$

where the elements of $\Pi \in \mathbb{R}^{m\times D}$ are drawn from a standard normal distribution, and the elements of $\mathbf{b} \in \mathbb{R}^m$ are drawn from the uniform distribution on $[0, 2\pi]$. A related but different random feature map was introduced in the context of Performers (53). There it was shown that instead of cosines, one can just as well use exponential functions

$$\phi(\mathbf{x}) = \sqrt{\frac{1}{m}}\exp\Big(\frac{-||\mathbf{x}||^2}{2}\Big)\exp(\Pi\mathbf{x}), \quad [14]$$

Note that due to the softmax normalization, any constant prefactors in Eq. **13** can be ignored (since they cancel in the numerator and denominator). If we assume an additional spherical normalization step before the random projection layer, so that all arguments to ϕ have constant norm, then the above activation functions may be written more plainly as

$$\phi(\mathbf{x}) = \cos(\Pi\mathbf{x} + \mathbf{b}) \quad \text{and} \quad \phi(\mathbf{x}) = \exp(\Pi\mathbf{x}).$$

Cosine tuning curves appear ubiquitously in neuroscience, across many different organisms (e.g., crickets, cats, rhesus monkeys) and many different brain areas (e.g., cerebellum, motor cortex, and hippocampus) (57, 58). The function $\exp(\cdot)$ is monotonic and positive, making it easy to implement from a biological perspective. For the exponential random feature function, the term $\exp(\frac{-||\mathbf{x}||^2}{2})$ may be interpreted as a homeostatic mechanism to ensure that firing rates do not become too large. We stress that while the aforementioned random feature maps are sufficient for approximating the softmax self-attention mechanism, there are infinitely many other activation functions that lead to valid (though potentially nonsoftmax) self-attention matrices.

General Case of Untied Weights

In this section, we relax the weight tying condition and generalize our construction to the case when $D \neq d$. While in the previous sections r acted as a gatekeeper for the weight matrix H, we will now *also* have r act as a gatekeeper for a few other weight matrices. Using the same variable names, consider the following neuron–astrocyte forward equations:

$$\begin{aligned} \mathbf{f} &= \mathbf{x} && \in \mathbb{R}^d \\ \mathbf{h} &= \phi\big[(1-r)W_K\mathbf{f} + rW_Q\mathbf{f}\big] && \in \mathbb{R}^m \\ \ell &= r(H \odot \tilde{P})\mathbf{h} + (1-r)W_V\mathbf{f} + r\mathbf{f} && \in \mathbb{R}^d, \end{aligned} \quad [15]$$

When $r = 0$, we recover the writing phase of Eq. **9**; when $r = 1$, we recover the reading phase equations of Eq. **11**. When we impose the weight tying constraint of $W_K = W_Q = W$ and $W_V = I$, we recover the original equations of Eq. **6**. Eq. **15** describes the neuron–astrocyte implementation of the general Transformer block without the weight sharing constraint imposed. The circuit diagram corresponding to Eq. **15** can be seen in Fig. 2*A*.

Numerical Validation

The results derived above have also been checked numerically. In Fig. 2*B*, one can see the error between the proposed neuron–astrocyte network and the actual AI Transformer block as a function of the ratio of the width of the hidden layer to the size of the token embedding. As expected from the theoretical analysis, the error between the two networks rapidly decreases as the hidden layer becomes wider. In practice, as the width of the hidden layer becomes 5 to 10 times the embedding dimension, the two networks produce very similar outputs. In Fig. 3*A*, we use the parameters of the ALBERT-base (59, 60) Transformer to generate a corresponding neuron–astrocyte model. In particular, we extracted the word embedding matrix, the encoder matrix, and the W_Q, W_K, W_V matrices from the first block of ALBERT-base. We then embedded and encoded the first 200 words of the abstract of this paper. We plugged these weights into two neuron–astrocyte networks Eq. **15**—one with $m = 10^3$ hidden neurons and one with $m = 10^5$ hidden neurons—and passed the tokens through the network. We extracted the astrocyte responses during the reading phase and plotted these along with the actual softmax

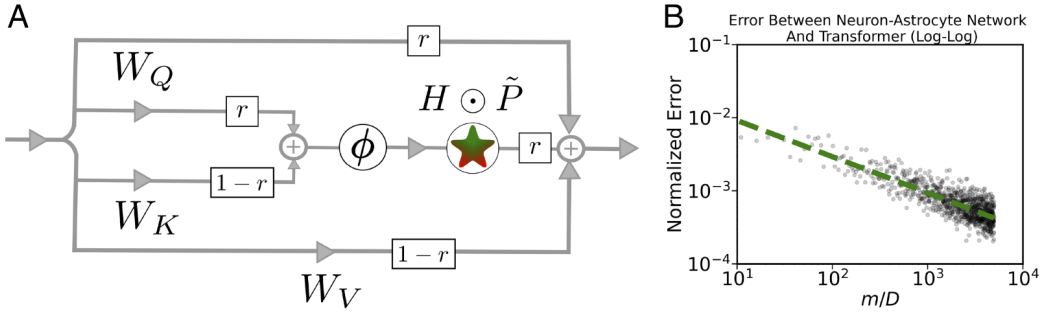

Fig. 2. (*A*) Circuit diagram of the full neuron–astrocyte model Eq. **15**, which implements a general (i.e., untied) Transformer block. (*B*) Error vs number of hidden units (m) in our network. As m increases, the difference between the output of the neuron–astrocyte circuit and the AI Transformer block decreases.

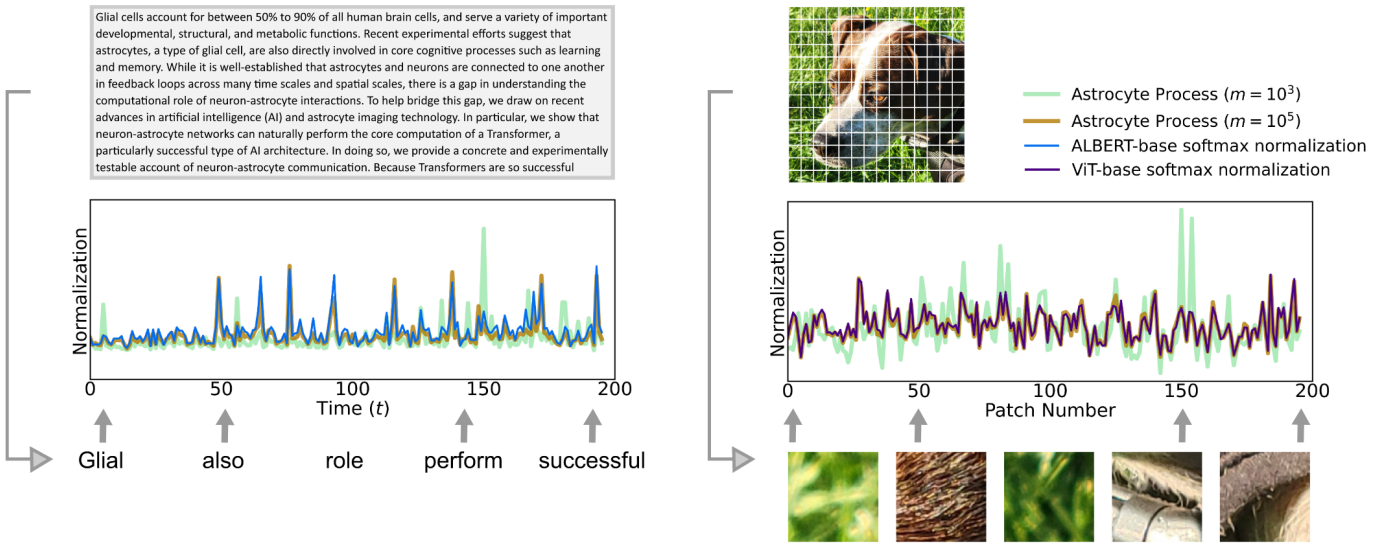

Fig. 3. (*Left*) Astrocyte traces for $m = 10^3$ and $m = 10^5$ compared against the "exact" softmax normalization terms of the ALBERT-base model. The tokens used were the first 200 words of the abstract of this paper. See "Numerical Validation" for details. (*Right*) Similar plot as in left side, but for a Vision Transformer. Instead of using embedded words as tokens, the model uses patches from an image.

normalization terms in ALBERT-base model. In Fig. 3*B*, we performed a similar "weight transfer" from a Vision Transformer model that was pretrained on ImageNet-21K (61, 62). In this case, the tokens were patches of an image, instead of words in a sentence. As expected from the theoretical derivation, for sufficiently large number of hidden units, neuron–astrocyte networks accurately describe computation performed by the Transformer models. The code to reproduce Fig. 3 is available in the following GitHub repository: https://github.com/kozleo/neuron-astrocyte-transformer.

Do We Need Astrocytes?

Although we are interested in addressing the scientific problem of how astrocytes participate in behavior, a natural question when positing any new brain mechanism is as follows: "Can the same behavior be achieved without this mechanism?" This section demonstrates that a Transformer circuit can *also* be constructed using neurons and bipartite synapses, together with a specialized divisive normalization achieved via shunting inhibition. The circuit is similar to Eq. **6**:

$$\begin{aligned} \mathbf{f} &= \mathbf{x} && \in \mathbb{R}^d \\ \mathbf{h} &= \phi(W\mathbf{f}) && \in \mathbb{R}^m \\ R &= \mathbf{g}^T\mathbf{h} && \in \mathbb{R} \\ \boldsymbol{\ell} &= \frac{r}{R}H\mathbf{h} + \mathbf{f} && \in \mathbb{R}^d. \end{aligned} \qquad [16]$$

The only difference between Eqs. **16** and **6** is the addition of a new element, R, and the removal of the astrocyte processes. Here, R is an inhibitory neuron that divisively normalizes feed-forward inputs into layer $\boldsymbol{\ell}$. However, it does not inhibit all feedfoward inputs equally. Despite both $\mathbf{h}$ and $\mathbf{f}$ being feed-forward inputs to layer $\boldsymbol{\ell}$, the divisive inhibition is only implemented on the inputs coming from layer $\mathbf{h}$. This can happen, for example, if the feed-forward synaptic inputs coming from layer $\mathbf{h}$ arrive at the dendritic tree close to where inhibitory inputs from neuron R shunt current flow, while the feed-forward inputs coming from layer $\mathbf{f}$ synapse far away from the shunting (63). Leaving the reading and writing phases untouched, circuit Eq. **16** implements the same forward pass as Eq. **6**.

While the proposed nonastrocytic circuit can, in theory, also implement a Transformer forward pass, it should be noted that there exists a controversy about the capability of shunting inhibition to implement divisive normalization (63, 64). Thus, the biologically plausibility of this circuit is questionable. Additionally—as we will discuss in the next section—the comparatively slower timescale of astrocytes provides a natural memory buffer when, e.g., accumulating and storing words in a sentence. Finally, it is possible that there are *many* ways to implement Transformers biologically, each with relative pros and cons. Different brain areas may implement Transformer-like computation using different circuitries. It is ultimately an experimental question to validate these theoretical hypothesis.

Timescales

One aspect of our model which we have yet to discuss is its timescale. Our circuit operates in two distinct phases: a reading phase and a writing phase. The reading phase does not involve any plasticity, so the only relevant timescale to compute is how long it takes to traverse the neuron–astrocyte-synapse pathway. Recent data indicate that astrocytes can sense and respond to neural activity on the order of a few hundreds of milliseconds (9, 65). The speed of the writing phase is limited by the speed of plasticity. There are two types of plasticity used in our model during the writing phase: 1) Hebbian plasticity between neurons and 2) presynaptic plasticity between neurons and astrocytic processes. In the case of neuron–neuron plasticity, there are experimental studies reporting a vast range of the relevant timescales. These include Hebbian plasticity (66–68), behavioral timescale plasticity (69–71), etc. The induction timescales for these plasticity mechanisms range from hundreds of milliseconds (70) to tens of minutes (67). In the case of STDP computational modeling studies, it is typically assumed that synaptic weights are adjusted instantaneously, by an amount proportional to the timing difference between pre-post synaptic spikes (72, 73). The neuron–astrocyte plasticity timescale is harder to establish, due to limitations in calcium recording technology. While fast calcium

transients in astrocyte processes have been recently recorded (9), and neuron–astrocyte plasticity has been experimentally observed (74), fast (e.g., <1 s) neuron–astrocyte plasticity has not been observed yet, possibly due to limitations of the calcium imaging technology.

Discussion

Here, we have built a computational neuron–astrocyte model which is functionally equivalent to an important AI architecture: the Transformer. This model serves a dual purpose. The first purpose is to provide a concrete, normative, computational account of how the communication between astrocytes and neurons subserves brain function. The second purpose is to provide a biologically plausible account of how Transformers might be implemented in the brain. While the feedback loop between neurons and astrocytes is well studied from an experimental perspective, there is comparatively little work studying it from the computational perspective (7). Astrocyte modeling studies tend to focus on either the biophysics of neuron-astrocyte or astrocyte signaling (75, 76) or the emergent computational properties of detailed neuron-astrocyte models (77–79). Fewer studies have focused on simpler, normative models of neuron–astrocyte networks (51, 80, 81).

An important feature of our model is that it is flexible enough to approximate any Transformer. In other words, we do not only show how to model a particular Transformer (i.e., one with weights that have already been trained for some specific task)—rather, we show how to approximate all *possible* Transformers using neurons and astrocytes. Given the demonstrated power and flexibility of Transformers, this generality can help to explain why astrocytes are so prevalent across disparate brain areas and species. Our model has several immediate implications. First, as calcium imaging technologies improve, it will become increasingly feasible to explicitly compare artificial representations in AI networks to representations in biological astrocyte networks—as is already done when comparing AI networks to biological neural networks (16, 22, 82). Given that astrocyte activity is thought to be tightly coupled to fMRI responses (83), natural language processing contexts such as (16) and (84) are already a promising place to look for astrocytic contributions to brain function. Additionally, we propose that our hypothesis could be refuted through studies involving targeted astrocyte manipulations. The brain's sensitivity to normal astrocyte function levels is evident. For instance, prior experimental studies have demonstrated that hippocampal astrocyte activation positively influences memory-related behaviors (85), whereas striatal astrocyte activation impairs attention (86). To challenge our hypothesis, we could train both a Transformer model and an animal subject to perform the same hippocampal-based memory task, such as one requiring path integration. Based on previous research, we anticipate a strong correlation between Transformer and hippocampal activations (87). If we could then selectively silence or modify hippocampal astrocytes in the animal subject and demonstrate that the representational similarity to the Transformer model remains unaffected, our hypothesis would be undermined. The main constraint of this approach lies in the present challenge of selectively inactivating astrocytes in a controlled and reversible fashion (1). Nevertheless, we anticipate that advancements in the field of astrocyte biology will eventually overcome these limitations.

Despite the exciting potential links between Transformers and the brain, it is worth noting that humans learn quite differently from Transformers. Transformers are extremely data-hungry, and consequently, training them requires a massive amount of energy (88). By contrast, the human brain runs on a smaller energy budget than a common laptop and does not require internet-scale training datasets to learn a language (89). In view of this fact, it may be more appropriate to view training a large Transformer as analogous to learning over evolutionary timescales, rather than the lifetime of a single individual (90).

Finally, a major roadblock in accepting Transformers as models of natural language processing (or, more generally, sequential processing) in the brain is that they require a memory buffer to store the tokens as they are presented. This is because the self-attention matrix is computed over all the tokens. Our paper proposes that neuron–astrocyte networks can perform this buffering naturally through spatial and temporal integration. Finally, and more speculatively, since astrocytes are implicated in many brain disorders and diseases, our work suggests that causal manipulations on Transformers can be used as a way to generate putative hypotheses for how astrocyte function goes astray in brain disorders and diseases (91, 92).

Data, Materials, and Software Availability. There are no data underlying this work. The code used in this work is available at GitHub repository (https://github.com/kozleo/neuron-astrocyte-transformer) (93).

ACKNOWLEDGMENTS. We thank Dan Gutfreund, John Hopfield, Martin Schrimpf, and Mriganka Sur for helpful comments and feedback. This work was completed while L.K. was an MIT-IBM Watson AI Lab Summer 2022 Intern. K.V.K. acknowledges funding from the following sources: BrightFocus Foundation Grant A2020833S, and National Institutes of Health Grant R01AG066171.

1. P. Kofuji, A. Araque, Astrocytes and behavior. *Annu. Rev. Neurosci.* **44**, 49–67 (2021).
2. B. L. Lind, A. R. Brazhe, S. B. Jessen, F. C. C. Tan, M. J. Lauritzen, Rapid stimulus-evoked astrocyte Ca^{2+} elevations and hemodynamic responses in mouse somatosensory cortex in vivo. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E4678–E4687 (2013).
3. A. Pinto-Duarte, A. J. Roberts, K. Ouyang, T. J. Sejnowski, Impairments in remote memory caused by the lack of Type 2 IP_3 receptors. *Glia* **67**, 1976–1989 (2019).
4. B. A. MacVicar, E. A. Newman, Astrocyte regulation of blood flow in the brain. *Cold Spring Harb. Perspect. Biol.* **7**, a020388 (2015).
5. W.-S. Chung, N. J. Allen, C. Eroglu, Astrocytes control synapse formation, function, and elimination. *Cold Spring Harb. Perspect. Biol.* **7**, a020370 (2015).
6. A. Kol, I. Goshen, The memory orchestra: The role of astrocytes and oligodendrocytes in parallel to neurons. *Curr. Opin. Neurobiol.* **67**, 131–137 (2021).
7. K. V. Kastanenka *et al*., A roadmap to integrate astrocytes into systems neuroscience. *Glia* **68**, 5–26 (2020).
8. M. López-Hidalgo, V. Kellner, J. Schummers, Astrocyte subdomains respond independently in vivo. bioRxiv [Preprint] (2019). https://doi.org/10.1101/675769 (Accessed 20 June 2019).
9. J. L. Stobart *et al*., Cortical circuit activity evokes rapid astrocyte calcium signals on a similar timescale to neurons. *Neuron* **98**, 726–735 (2018).
10. M. Yu *et al*., Glia accumulate evidence that actions are futile and suppress unsuccessful behavior. *Cell* **178**, 27–43 (2019).
11. J. Nagai *et al*., Behaviorally consequential astrocytic regulation of neural circuits. *Neuron* **109**, 576–596 (2021).
12. Z. Lin *et al*., Entrainment of astrocytic and neuronal Ca^{2+} population dynamics during information processing of working memory in mice. *Neurosci. Bull.* **38**, 474–488 (2022).
13. A. Vaswani *et al*., "Attention is all you need" in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017), vol. 30. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (Accessed 6 December 2017).
14. M. Toneva, L. Wehbe, Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Adv. Neural. Inf. Process. Syst.* **32**, 14954–14964 (2019).
15. D. Krotov, J. J. Hopfield, "Large associative memory problem in neurobiology and machine learning" in *International Conference on Learning Representations* (OpenReview.net, 2021).
16. M. Schrimpf *et al*., The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2105646118 (2021).
17. D. Tyulmankov, C. Fang, A. Vadaparty, G. R. Yang, "Biological learning in key-value memory networks" in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2021), vol. 34, pp. 22247–22258.
18. T. Bricken, C. Pehlevan, Attention approximates sparse distributed memory. *Adv. Neural. Inf. Process. Syst.* **34**, 15301–15315 (2021).

19. J. C. R. Whittington, J. Warren, T. E. J. Behrens, Relating transformers to models and neural representations of the hippocampal formation. arXiv [Preprint] (2022). http://arxiv.org/abs/2112.04035 (Accessed 15 March 2022).
20. C. Caucheteux, J.-R. King, Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 1–10 (2022).
21. T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers. arXiv [Preprint] (2021). http://arxiv.org/abs/2106.04554 (Accessed 15 June 2021).
22. D. L. K. Yamins *et al.*, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
23. J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554–2558 (1982).
24. G. Perea, M. Navarrete, A. Araque, Tripartite synapses: Astrocytes process and control synaptic information. *Trends Neurosci.* **32**, 421–431 (2009).
25. C. S. Von Bartheld, J. Bahney, S. Herculano-Houzel, The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *J. Comp. Neurol.* **524**, 3865–3895 (2016).
26. M. M. Halassa, T. Fellin, H. Takano, J.-H. Dong, P. G. Haydon, Synaptic islands defined by the territory of a single astrocyte. *J. Neurosci.* **27**, 6473–6477 (2007).
27. N. A. Oberheim *et al.*, Uniquely hominid features of adult human astrocytes. *J. Neurosci.* **29**, 3276–3287 (2009).
28. A. Semyanov, A. Verkhratsky, Astrocytic processes: From tripartite synapses to the active milieu. *Trends Neurosci.* **44**, 781–792 (2021).
29. A. Verkhratsky, A. Butt, *Glial Neurobiology: A Textbook* (John Wiley & Sons, 2007).
30. E. A. Newman, New roles for astrocytes: Regulation of synaptic transmission. *Trends Neurosci.* **26**, 536–542 (2003).
31. G. R. J. Gordon *et al.*, Astrocyte-mediated distributed plasticity at hypothalamic glutamate synapses. *Neuron* **64**, 391–403 (2009).
32. J.-Y. Sul, G. Orosz, R. S. Givens, P. G. Haydon, Astrocytic connectivity in the hippocampus. *Neuron Glia Biol.* **1**, 3–11 (2004).
33. N. Kuga, T. Sasaki, Y. Takahara, N. Matsuki, Y. Ikegaya, Large-scale calcium waves traveling through astrocytic networks in vivo. *J. Neurosci.* **31**, 2607–2614 (2011).
34. E. Scemes, C. Giaume, Astrocyte calcium waves: What they are and what they do. *Glia* **54**, 716–725 (2006).
35. R. Bommasani *et al.*, On the opportunities and risks of foundation models. arXiv [Preprint] (2021). http://arxiv.org/abs/2108.07258 (Accessed 12 June 2022).
36. T. Brown *et al.*, Language models are few-shot learners. *Adv. Neural. Inf. Process. Syst.* **33**, 1877–1901 (2020).
37. OpenAI, ChatGPT: Optimizing language models for dialogue (2022). https://openai.com/blog/chatgpt/ (Accessed 12 May 2022).
38. A. Dosovitskiy *et al.*, An image is worth 16x16 words: Transformers for image recognition at scale. arXiv [Preprint] (2020). http://arxiv.org/abs/2010.11929 (Accessed 3 June 2021).
39. S. Hochreiter, *Untersuchungen zu Dynamischen Neuronalen Netzen* (Diploma, Technische Universität München, 1991), vol. 91.
40. Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Networks* **5**, 157–166 (1994).
41. R. Pascanu, T. Mikolov, Y. Bengio, "On the difficulty of training recurrent neural networks" in *International Conference on Machine Learning* (PMLR, 2013), pp. 1310–1318.
42. M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision transformers see like convolutional neural networks? *Adv. Neural. Inf. Process. Syst.* **34**, 12116–12128 (2021).
43. J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization. arXiv [Preprint] (2016). http://arxiv.org/abs/1607.06450 (Accessed 21 June 2016).
44. Y. Shen, J. Wang, S. Navlakha, A correspondence between normalization strategies in artificial and biological neural networks. *Neural Comput.* **33**, 3179–3203 (2021).
45. M. E. Sander, P. Ablin, M. Blondel, G. Peyré, "Sinkformers: Transformers with doubly stochastic attention" in *International Conference on Artificial Intelligence and Statistics* (PMLR, 2022), pp. 3515–3530.
46. Y. Yang, Z. Huang, D. Wipf, Transformers from an optimization perspective. arXiv [Preprint] (2022). http://arxiv.org/abs/2205.13891 (Accessed 27 May 2022).
47. P. Kanerva, *Sparse Distributed Memory* (MIT Press, 1988).
48. G. Perea, A. Araque, Astrocytes potentiate transmitter release at single hippocampal synapses. *Science* **317**, 1083–1086 (2007).
49. G. Perea, M. Navarrete, A. Araque, Tripartite synapses: Astrocytes process and control synaptic information. *Trends Neurosci.* **32**, 421–431 (2009).
50. M. De Pittà, N. Brunel, A. Volterra, Astrocytes: Orchestrating synaptic plasticity? *Neuroscience* **323**, 43–61 (2016).
51. V. Ivanov, K. Michmizos, Increasing liquid state machine performance with edge-of-chaos dynamics organized by astrocyte-modulated plasticity. *Adv. Neural. Inf. Process. Syst.* **34**, 25703–25719 (2021).
52. A. Rahimi, B. Recht, Random features for large-scale kernel machines. *Adv. Neural. Inf. Process. Syst.* **20** (2007).
53. K. Choromanski *et al.*, Rethinking attention with performers. arXiv [Preprint] (2020). http://arxiv.org/abs/2009.14794 (Accessed 30 September 2020).
54. H. Peng *et al.*, Random feature attention. arXiv [Preprint] (2021). http://arxiv.org/abs/2103.02143 (Accessed 19 March 2021).
55. A. Katharopoulos, A. Vyas, N. Pappas, F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention" in *International Conference on Machine Learning* (PMLR, 2020), pp. 5156–5165.
56. D. D. Rasmusson, The role of acetylcholine in cortical synaptic plasticity. *Behav. Brain Res.* **115**, 205–218 (2000).
57. A. P. Georgopoulos, J. F. Kalaska, R. Caminiti, J. T. Massey, On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J. Neurosci.* **2**, 1527–1537 (1982).
58. E. Salinas, L. F. Abbott, Vector reconstruction from firing rates. *J. Comput. Neurosci.* **1**, 89–107 (1994).
59. Z. Lan *et al.*, ALBERT: A Lite BERT for self-supervised learning of language representations. arXiv [Preprint] (2019). http://arxiv.org/abs/1909.11942 (Accessed 9 February 2020).
60. T. Wolf *et al.*, Huggingface's transformers: State-of-the-art natural language processing. arXiv [Preprint] (2019). http://arxiv.org/abs/1910.03771 (Accessed 14 July 2020).
61. B. Wu *et al.*, Visual transformers: Token-based image representation and processing for computer vision. arXiv [Preprint] (2020). https://doi.org/10.48550/arXiv.2006.03677 (Accessed 20 November 2020).
62. J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database" in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), pp. 248–255.
63. F. S. Chance, L. F. Abbott, Divisive inhibition in recurrent networks. *Netw. Comput. Neural Syst.* **11**, 119 (2000).
64. G. R. Holt, C. Koch, Shunting inhibition does not have a divisive effect on firing rates. *Neural Comput.* **9**, 1001–1013 (1997).
65. A. Semyanov, C. Henneberger, A. Agarwal, Making sense of astrocytic calcium signals–From acquisition to interpretation. *Nat. Rev. Neurosci.* **21**, 551–564 (2020).
66. H. Markram, J. Lübke, M. Frotscher, B. Sakmann, Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **275**, 213–215 (1997).
67. G. Bi, M. Poo, Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* **18**, 10464–10472 (1998).
68. M. A. Erickson, L. A. Maramara, J. Lisman, A single brief burst induces GluR1-dependent associative short-term potentiation: A potential mechanism for short-term memory. *J. Cogn. Neurosci.* **22**, 2530–2540 (2010).
69. K. C. Bittner, A. D. Milstein, C. Grienberger, S. Romani, J. C. Magee, Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science* **357**, 1033–1036 (2017).
70. J. C. Magee, C. Grienberger, Synaptic plasticity forms and functions. *Annu. Rev. Neurosci.* **43**, 95–117 (2020).
71. L. Z. Fan *et al.*, All-optical physiology resolves a synaptic basis for behavioral timescale plasticity. *Cell* **186**, 543–559.e19 (2023).
72. S. Song, K. D. Miller, L. F. Abbott, Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nat. Neurosci.* **3**, 919–926 (2000).
73. J. Sjöström *et al.*, Spike-timing dependent plasticity. *Scholarpedia* **35**, 1362 (2010).
74. W. Croft, K. L. Dobson, T. C. Bellamy, "Equipping glia for long-term integration of network activity" in *Neural Plasticity, Plasticity of Neuron-Glial Transmission* (Hindawi, 2015).
75. A. Witthoft, G. E. Karniadakis, A bidirectional model for communication in the neurovascular unit. *J. Theor. Biol.* **311**, 80–93 (2012).
76. L. P. Savtchenko, D. A. Rusakov, Regulation of rhythm genesis by volume-limited, astroglia-like signals in neural networks. *Philos. Trans. R. Soc. B: Biol. Sci.* **369**, 20130614 (2014).
77. M. De Pittà, N. Brunel, Multiple forms of working memory emerge from synapse–astrocyte interactions in a neuron–glia network model. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2207912119 (2022).
78. S. Becker, A. Nold, T. Tchumatchenko, Modulation of working memory duration by synaptic and astrocytic mechanisms. *PLoS Comput. Biol.* **18**, e1010543 (2022).
79. S. Y. Gordleeva *et al.*, Modeling working memory in a spiking neuron network accompanied by astrocytes. *Front. Cell. Neurosci.* **15**, 631485 (2021).
80. G. Tang, I. E. Polykretis, V. A. Ivanov, A. Shah, K. P. Michmizos, "Introducing astrocytes on a neuromorphic processor: Synchronization, local plasticity and edge of chaos" in *Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop* (Association for Computing Machinery, New York, NY, 2019), pp. 1–9.
81. E. J. Peterson, What can astrocytes compute? bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.10.20.465192 (Accessed 1 December 2022).
82. M. Schrimpf *et al.*, Brain-score: Which artificial neural network for object recognition is most brain-like? bioRxiv [Preprint] (2020). https://doi.org/10.1101/407007 (Accessed 9 May 2018).
83. C. R. Figley, P. W. Stroman, The role(s) of astrocytes and astrocyte activity in neurometabolism, neurovascular coupling, and the production of functional neuroimaging signals. *Eur. J. Neurosci.* **33**, 577–588 (2011).
84. S. Kumar *et al.*, Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.06.08.495348 (Accessed 9 May 2018).
85. A. Adamsky, I. Goshen, Astrocytes in memory function: Pioneering findings and future directions. *Neuroscience* **370**, 14–26 (2018).
86. J. Nagai *et al.*, Hyperactivity with disrupted attention by activation of an astrocyte synaptogenic cue. *Cell* **177**, 1280–1292 (2019).
87. J. C. R. Whittington, J. Warren, T. E. J. Behrens, Relating transformers to models and neural representations of the hippocampal formation. arXiv [Preprint] (2021). http://arxiv.org/abs/2112.04035 (Accessed 15 March 2022).
88. D. Patterson *et al.*, Carbon emissions and large neural network training. arXiv [Preprint] (2021). http://arxiv.org/abs/2104.10350 (Accessed 21 May 2021).
89. V. Balasubramanian, Brain power. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2107022118 (2021).
90. F. Geiger, M. Schrimpf, T. Marques, J. J. DiCarlo, Wiring up vision: Minimizing supervised synaptic updates needed to produce a primate ventral stream. bioRxiv [Preprint] 2020. https://doi.org/10.1101/2020.06.08.140111 (Accessed 6 August 2020).
91. C. Escartin *et al.*, Reactive astrocyte nomenclature, definitions, and future directions. *Nat. Neurosci.* **24**, 312–325 (2021).
92. V. Volman, M. Bazhenov, T. J. Sejnowski, Computational models of neuron–astrocyte interaction in epilepsy. *Front. Comput. Neurosci.* **6**, 58 (2012).
93. L. Kozachkov, D. Krotov, Building Transformers from Neurons and Astrocytes. *GitHub*. https://github.com/kozleo/neuron-astrocyte-transformer. Deposited 15 February 2023.