Astrocytes: the Transformers of the Brain

ME 225NN, Winter 2025

Raaghav Thirumaligai

June 8, 2025

Contents

1 Introduction

1.1 Problem Description and Motivation

The transformer architecture has redefined machine learning across various domains, including natural language processing, molecule generation, and high-dimensional image processing. The transformer's success stems largely from its self-attention mechanism and inherent parallelizability, enabling efficient learning of complex data representations. [1–4] Given its broad applicability driven by a relatively simple mechanism, there is a strong interest in understanding how this computational mechanism might have a biological analogy. This motivation leads to exploring neuron-astrocyte interactions as a promising biological representation of transformer computations. While astrocytes and their feedback mechanisms have been studied, the literature lacks computational view of neuron - astrocyte communication. In their paper, *Building transformers from neurons and astrocytes*, Kozachkov, Kastanenka and Krotov address the gap by constructing a model for how astrocytes may interact with neurons to represent computations analogous to transformers.

1.2 Literature Review

Transformers, introduced by a team of researchers from Google in 2017 [2] in the revolutionary paper *Attention is All You Need*, rely fundamentally on self-attention to learn global dependencies efficiently in data sequences. Recent advancements have extended their success beyond language to domains such as molecular modeling (Smaldone et al.(2025) [3]) and image processing (Porcu et al. (2024) [4]). Despite their computational success, biological parallels to transformers remain understudied. Kozachkov et al. (2023) [1] propose neuron-astrocyte networks as biological analogs for transformers, highlighting computational gaps in understanding neuron-astrocyte communication. Newman et al. (2003) and Perea et al. (2009) [5, 6] have characterized the functional unit of neuron-astrocyte interactions as the tripartite synapse, comprising the presynaptic neuron, postsynaptic neuron, and an astrocyte process; emphasizing that to understand synaptic function, the tripartite synapse must be considered. Araque et al. (1999) [7] further demonstrated that astrocytes, upon detecting neuronal activity, can modulate synaptic strength via calcium signaling, forming essential feedback loops. A single astrocyte has a region of influence where it can monitor and modify neuronal behavior. As we have seen across machine learning, simply increasing the scale of a model in parameters or layers often leads to increased performance. [8] Therefore, it is interesting when we consider that rodents have astrocytes that cover $\sim$ 20,000-120,000 synapses while human astrocytes can reach $\sim$ 270,000 - 2,000,000 synapses,

$$\begin{aligned} m_{rodent} &\in [20000, 120000] \\ m_{human} &\in [270000, 2000000] \end{aligned} \tag{1}$$

leading to speculation by Oberheim et al. (2009) [9], that the size of this region is correlated with the computational power and that the coupling of glial cell / neuron processing may increase intelligence.

1.3 Summary of Report Content

In this review, we first cover some background on astrocytes/ neurons and cover the fundamental transformer architecture. Then, we examine the framework proposed by Kozachkov et al., highlighting how astrocyte mediated feedback could biologically implement self-attention mechanisms

(which are fundamental to the transformer). Finally, we discuss implications of scaling neuron-astrocyte networks, considering the broader relevance to computational neuroscience and potential future research directions.

2 Preliminaries and Notation

In order to explore how astrocytes could perform the role of a biological transformer, we need to cover the background relevant to this. The broad neuron information primarily comes from *Campbell Biology* in *Chapter 48: Neurons, Synapses, and Signaling* unless otherwise specified. [10]

2.1 Biological Neuron

The biological neuron is a cell which takes stimuli as an input and outputs a fixed shape pulse. Varying these stimuli will result in a change in the frequency of the pulse. Neurons are excitable cells that encode information via action potentials. At rest, the membrane potential is approximately

$$V_{\text{rest}} \approx -70 \text{ mV},$$

maintained by ionic gradients (Na^+, K^+) and active sodium and potassium pumps. When sufficiently depolarized due to a strong stimulus (crossing a threshold $V_{\text{th}} \approx -55$ mV), voltage gated Na^+ channels open, causing a rapid upswing of membrane voltage up to about +50 mV, followed by repolarization and a brief undershoot before returning to rest. The timing and frequency of these spikes carry information downstream. The shape of this pulse is shown in Figure. 1. Suppose a neuron generates a signal from a stimulus. As the signal travels outwards, it moves to the furthest edges of the cell on long tubes called the axons. At the very end, the information must be transmitted from one neuron to the next. This is done at a junction called the synapse. The synapse allows for an electrical signal to be converted into neurotransmitters which are able to travel between cells. When these neurotransmitters leave the synapse, the next cell senses this stimulus and its own signal begins. Neurons additionally have plasticity which is their ability to adapt due to activity and change over time. This means that given the exact same stimulus, the same neuron after having different experiences could give a different response. This is a fundamentally important idea for the transformer representation.

2.2 Astrocyte

Seeing how capable and intricate this neuron is, it may beg the question of what the point of the astrocyte is. Addressing this first before introducing the astrocyte; to construct a biological model for a transformer, one must require that neurons are able to apply a divisive normalization which seems to be disputed in other literature. [1] Taking this at face value, models are of questionable plausibility without the astrocyte, motivating studying what they are.

The astrocyte lives in an arrangement called the tripartite synapse along with two other neurons (the presynaptic and post synaptic neuron). Most of the synapses in the brain are tripartite, meaning that our model earlier discussing the synapse alone was incomplete. Astrocytes are capable of sensing the neuronal activity and will respond by increasing their intracellular calcium concentration. In neurons, this releases neurotransmitters creates a feedback loop with the neurons. The astrocyte response can lead to divisive normalization much like the scaling step in transformer self-attention [1].

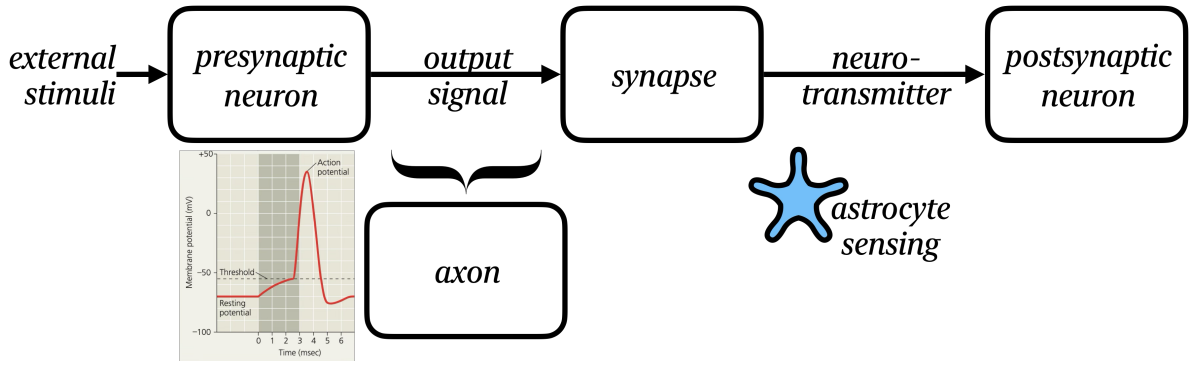

Figure 1: A schematic for a signal travelling through a neuron. The external stimuli will cause the neuron's internal potential to become less negative and once a threshold is crossed, a wave of action potential travels through the axon towards the synapse. At the synapse, an astrocyte sits and observes and perhaps modifies behavior. The synapse releases neurotransmitters which diffuse towards the postsynaptic neuron that senses this as a stimulus. (self-gen except AP plot from [10])

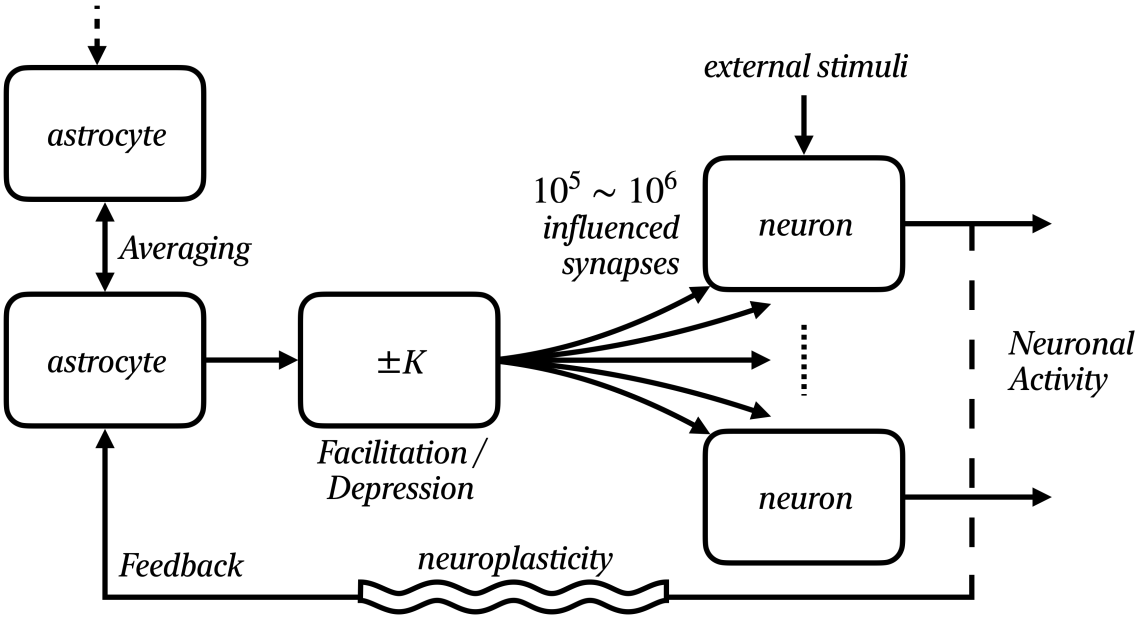

Figure 2: Tripartite Synapse. This is a block diagram I made to describe the processes/mechanisms occurring within the system that are relevant to its representation as a biological transformer. There are large scale networks of astrocytes which have spatial averaging of their calcium levels. [1, 11] Upon activation, the astrocytes' calcium levels increase and modulate the neurons with a common proportional scaling by releasing neurotransmitters (when considering glutamate synapses). [5, 12] Upon neuronal activity, astrocytes are stimulated through a plastic signaling pathway (meaning the feedback strength can change), triggering a calcium increase within the astrocyte. (self-gen fig)

2.3 Transformer Introduction

The transformer is a neural architecture which allows for much greater parallelization than other machine learning methods such as recurrent neural networks. While a recurrent neural network will process inputs sequentially, the transformer has access to all previous inputs through the self-

attention mechanism. [1] This allows the transformer to develop some 'understanding' of what an input means. To elaborate on this more, consider Figure. 3, where an input is taken and processed through the transformer architecture. The input can be any form of data which goes through some conversion to generate a vector that represents a small element of the data and some information about its location. This could be a symbol or piece of a word for a text input or the values of set of pixels in an image. Once the data is embedded into a vector, if the input token is the same then the vector should be the same. For example, if we have the following sentences:

1. The Imaginary Axis is where I put my ***poles***
2. I can't ski without my ***poles***

Upon embedding, the tokens representing ***poles*** are identical.[1] Clearly, the first sentence comes from a dubious controls engineer, but alarm bells should be going off in your *human* brain since you have intuition that these are very different poles! This additional meaning that our brains develop for the token from just the context is what the self-attention mechanism will attempt to recover. It is specifically here where biological interpretations are lacking and the focus where the astrocytes could play a role. [1] There is an important piece of the transformer that comes after this within a feed-forward block often called the multi-layer perceptron. It is thought that the weights here may be where transformers store facts that are not directly in the input data. [8] Once the tokens have been processed in the transformer block enough times to encode some meaning, it is passed to a linear layer that will generate the next token. The output is then normalized with some temperature to generate a more random and natural guess for the next token. [13] This normalization ($\text{soft} - \arg - \max$) is important and should be understood before diving into the self-attention mechanism.

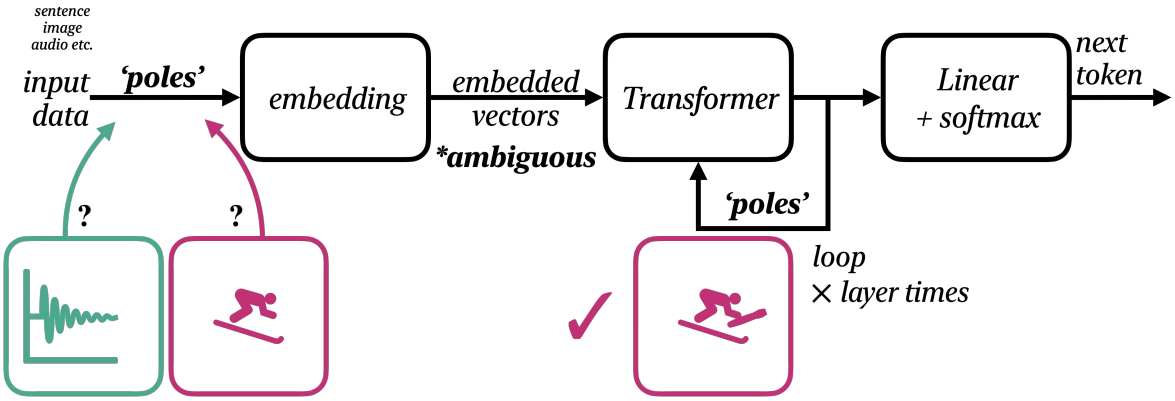

Figure 3: Illustration of how a transformer 'understands' an initially ambiguous token. Both an underdamped response (green) and a skiing scene (pink) are mapped to the same token "poles" and thus share identical embeddings (left). These are broad ideas from the context we hope to capture. After passing through the stacked self-attention and feed-forward layers within the transformer block enough times, the network resolves the correct skiing-poles interpretation (right). This is then used when generating the next token from the vector representing ***poles*** that has been augmented to encode some understanding since it is the final vector in the input. (self-gen fig)

[1]Technically they need to be in the same location in the sentence, but this is not super important for the point I'm making.

2.4 Softmax

The softmax function is used in the Self-Attention mechanism. It is a normalization function which takes a vector and exponentiates element wise then divides each element by the sum of the exponentiated elements of the vector:

$$\forall x \in \mathbb{R}^n, \quad \text{softmax}_T(x) = \frac{1}{\sum_{i=1}^{n} \exp\left(\frac{x_i}{T}\right)} \begin{bmatrix} \exp\left(\frac{x_1}{T}\right) \\ \vdots \\ \exp\left(\frac{x_n}{T}\right) \end{bmatrix}. \tag{2}$$

Tuning the temperature parameter can be seen in Figure. 4. This acts as an averaging parameter but it accentuates the maximal element when $T \to 0$. This function requires a biologically plausible representation which astrocytes can provide through their averaging behavior.

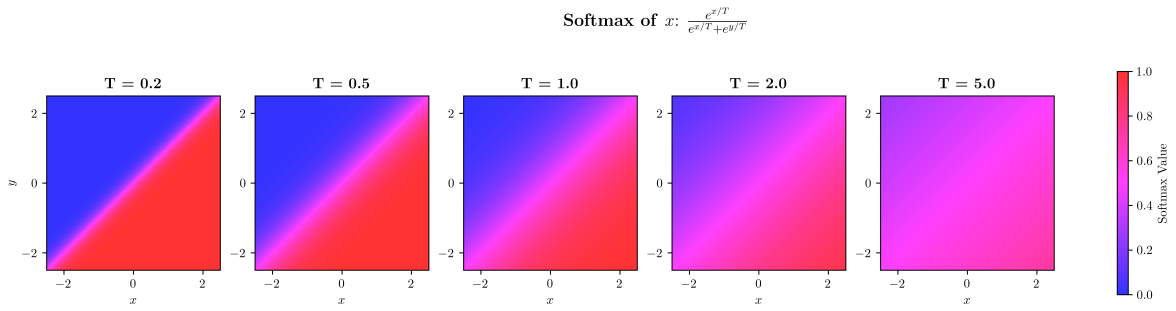

Figure 4: This figure illustrates how the 2D softmax changes when temperature is tuned. When the temperature approaches zero, we have a straightforward maximum function where value = 1 is assigned to the maximum element. As the temperature is adjusted to be greater or less than 1, we observe either sharper or softer behavior in the normalization. Increasing the temperature increases the likelihood of generating less probable tokens. Finding the optimal temperature is crucial to transition from robotic to natural language generation without producing nonsensical output. (self-gen fig)

2.5 Self-Attention

A visual representation of this mechanism is given in Figure. 5 which takes a peek into the transformer, but more clarity behind the computation comes from the math. To give some data to the transformer, it must first be embedded as we discussed earlier. Once the data is embedded, each token can be combined into a matrix. The input given to the transformer is represented by the matrix,

$$X \in \mathbb{R}^{d \times N}, \tag{3}$$

where d is the length of each token and N is the number of tokens. Within the transformer we generate some matrices known as the key, value, and query:

$$\mathbf{q}_t = W_Q \, \mathbf{x}_t, \qquad Q = W_Q \, X. \tag{4}$$

$$\mathbf{k}_t = W_K \, \mathbf{x}_t, \qquad K = W_K \, X, \tag{5}$$

$$\mathbf{v}_t = W_V \, \mathbf{x}_t, \qquad V = W_V \, X, \tag{6}$$

where,

$$W_{Q,K} \in \mathbb{R}^{D \times d},$$
$$W_V \in \mathbb{R}^{d \times d},$$

and typically, $D << d$. Each of these vectors is generated from an embedded token from the input data and equivalently the matrix representing the data can be multiplied to do the same. The weights of these matrices are tuned commonly using back propagation algorithms. The query and key vectors are intended to represent our token in smaller dimensional spaces. The alignment between these vectors represent some correlation of meaning. We look at how all of our key and query vectors align with each other. If a key and query are aligned closely, that means our embedded token for the query should have some directionality more along the token representing the key. This alignment is measured by the dot product between the vectors and this can effectively be done with the matrices then normalizing to identify which correlations are most significant for each query. [2]

$$\text{softmax}(K^T Q) \tag{7}$$

To find how each token should be influenced by this normalized alignment, we multiply by the value vector, represented in the initial embedding space.

$$\text{SelfAttention} = V\text{softmax}(K^T Q) \tag{8}$$

This is the self-attention mechanism. (Figure. 5)

2.6 Transformer

The Self-Attention mechanism is main novelty introduced and the key component that lacks biological representation. Yet, the transformer consists of other components. Within the transformer, the augmented value is then summed with the initial tokens to adjust the embedding to represent the initial data appropriately with the relevant correlations, hopefully including a more 'meaningful' token.

$$X + V\text{softmax}(K^T Q) \tag{9}$$

A LayerNorm is applied then it enters a feed-forward network, summed with a feed through term. This has a LayerNorm applied once more and completes one loop through the transformer.

$$Y = \text{LayerNorm}(X + V\text{softmax}(K^T Q)) \tag{10}$$
$$\text{Transformer}(X) = \text{LayerNorm}(\text{FFN}(Y) + Y) \tag{11}$$

It is important to note that this is known as a single head of attention. A multi-headed attention mechanism will do the same steps but within the attention blocks, the operations occur many times and are done with various weights for key, query, and value. [2]

[2]The sense of position is important here since the vectors from the 'future' should not influence the vectors in the past. This means that the tokens in a sentence should only be considered against the tokens prior to them. This is done in practice by setting the non-causal key/query products to negative ∞ before taking the softmax of them. [8]

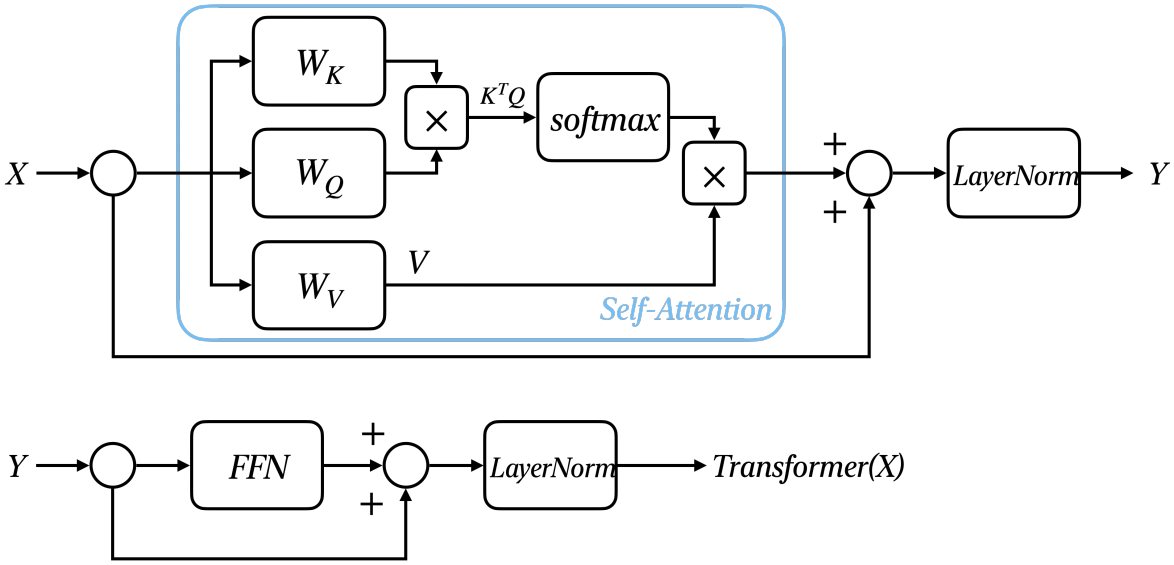

Figure 5: The block diagram representing the transformer is shown. Our data enters as tokens within the matrix X. It is scaled by the appropriate weights and the softmax normalizes the query-key alignment. This is multiplied by the value, V, which gives a sense for which features are in need of realignment. This is summed with feed through tokens, so they are augmented into the proper direction. The transformer applies a LayerNorm, a Feed Forward Network (sometimes a multi-layer perceptron) and LayerNorm to give one loop of the transformer. (self-gen fig)

2.7 Notation

Below are the key variables and matrices **used in the neuron-astrocyte model**, particularly as they appear in the forward equations:

- d is the dimension of the input token.
- $\mathbf{f} = \mathbf{x} \in \mathbb{R}^d$: the input feature vector representing a token in a sequence, of dimension d.
- m is the number of units in the hidden layer.
- $W_K, W_Q \in \mathbb{R}^{m \times d}$: learnable weight matrices corresponding to key and query transformations, respectively.
- $W_V \in \mathbb{R}^{d \times d}$: learnable weight matrix corresponding to value transformation.
- $\phi(\cdot)$: an approximate feature map ($\sim$ exponential dot product) used to approximate softmax style attention in a biologically plausible manner.
- $r \in [0, 1]$: state controlling reading($r = 1$) vs writing($r = 0$) phase of the model.
- $\mathbf{h} \in \mathbb{R}^m$: an intermediate hidden representation derived from projecting $\mathbf{f}$ through a combination of W_K and W_Q, followed by the activation.
- $H \in \mathbb{R}^{d \times m}$: a learnable matrix encoding how attention is modulated by astrocyte activity.
- $\tilde{P} \in \mathbb{R}^{d \times m}$: an astrocyte - derived matrix encoding normalized synaptic interactions, functioning analogously to the softmax attention matrix.

- $\odot$: element-wise (Hadamard) product.
- $\boldsymbol{\ell} \in \mathbb{R}^d$: the final output of the layer, combining astrocytic modulation, value projections, and residual input.

This notation encodes a biologically inspired approximation of the Transformer self-attention mechanism, where astrocyte modulation replaces softmax normalization and synaptic weight updates are locally computed.

3 The Neuron-Astrocyte Model

The spatial mechanic of the self-attention mechanism which allows for longer range dependencies is similar to the idea of astrocytes, sharing information with each other through large scale averaging dynamics. Kozachkov et al. propose the following neuron-astrocyte forward equations:

$$\begin{aligned} \mathbf{f} &= \mathbf{x} && \in \mathbb{R}^d \\ \mathbf{h} &= \phi\big[(1-r)W_K\mathbf{f} + rW_Q\mathbf{f}\big] && \in \mathbb{R}^m \\ \boldsymbol{\ell} &= r(H \odot \tilde{P})\mathbf{h} + (1-r)W_V\mathbf{f} + r\mathbf{f} && \in \mathbb{R}^d, \end{aligned} \tag{12}$$

where f is the first layer, h is the hidden layer and l is the last layer. This is represented in a block diagram in Figure. 6

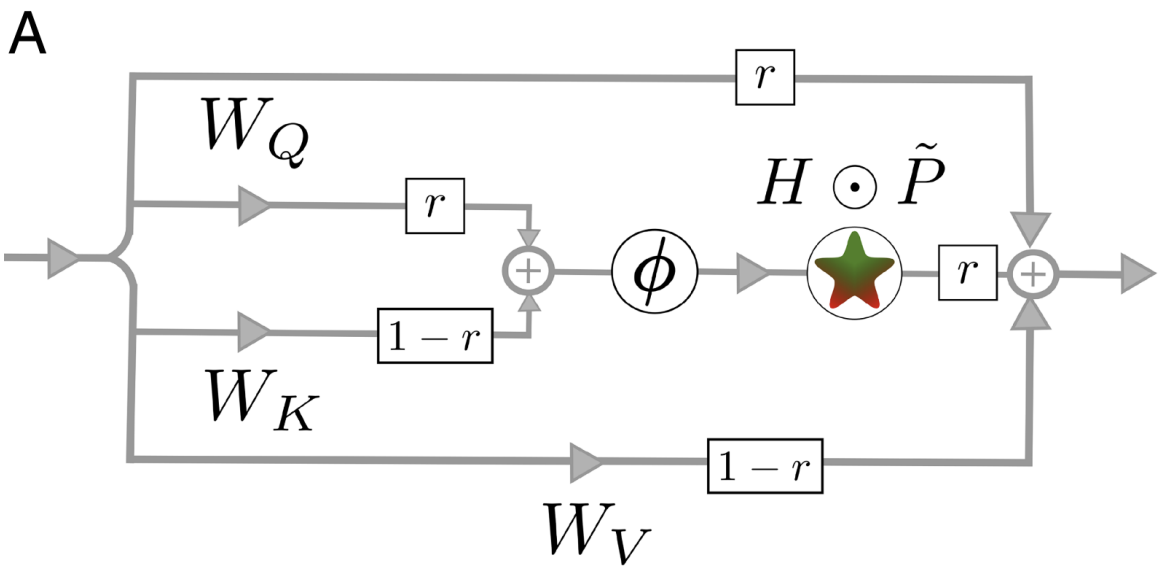

Figure 6: The block diagram representing the system of equation (12). The tokens are input(left) and scaled by according weights. The element wise multiplication of $\tilde{P}$ represents the scaling by astrocyte activity. Depending on the reading or writing phase, r will allow or inhibit the hidden layer from being expressed. When writing($r = 0$), the output simply becomes the value of the token. (not self-gen, fig from [1])

Going through equation (12), the first layer just takes the input token, and tokens stream through the system. In the hidden layer, the neural activation function ϕ acts on the weighted key and query. There is a writing phase that occurs, representing some possible neuromodulator

released. Effectively this allows us to update weights with some plasticity. The key seems to never be processed actually since if it is in the reading phase, it is multiplied by $(1-r)|_{r=1}$ and in the writing phase, the hidden layer is multiplied by 0 in the last layer equation. However, the keys are actually used to generate the $\tilde{P}$ matrix which is used with the element wise product against H. All of this is then summed with the initial token in the reading phase.

Within the writing phase, $\tilde{P}$ is some plasticity in our system and consists of the neuron to neuron plasticity through the activation function acting on the queries. There is neuron to astrocyte plasticity as well which is represented through the keys in the activation function. The plasticity looks at the neural activation of the query of the t^{th} token and see how it aligns with the neural activation of each prior key, then sum and invert because we of the inhibition capacity for astrocytes.

$$\tilde{P} := \frac{m}{\sum_{j=1}^{N} \phi(k_j)^T \cdot \phi(q_t)}. \tag{13}$$

This matrix represents the astrocyte dynamics. The assumption is that averaging occurs across astrocytes so that each element of P is the same. This gets multiplied by H defined as:

$$H := \tfrac{1}{m} V \phi(K)^T. \tag{14}$$

This is a weight matrix updated by Hebbian plasticity and the learning rate is $\frac{1}{m}$.

There are many analogies between this and the transformer. The first token is getting augmented in the reading phase by some weights depending on the key and queries. This behavior is allowing each token to be influenced by each other token just like the self attention block. Within the reading phase, the number of our hidden neurons is defined by m. This is equivalent to the region of influence that our astrocytes have which is suspected to be related to intelligence. [9] When we have the reading phase, there is a time delay that is occurs similar to the response delay when comparing the calcium diffusion to the timescale of a neuron firing.

4 Simulation

In the paper [1], the numerical validation is done by taking embedding matrix, encoder matrix and the key, query, and value weights from ALBERT-base. A comparison is done to see how each token is normalized by the neuron-astrocyte model versus the softmax for ALBERT. The plot they show seems to imply that increasing the number of neurons gets closer to the softmax behavior. In Figure. 7, I have used the code provided by [1] and recreated their **Fig.3** using the average of the values representing rodents and humans' astrocyte influence in (1).

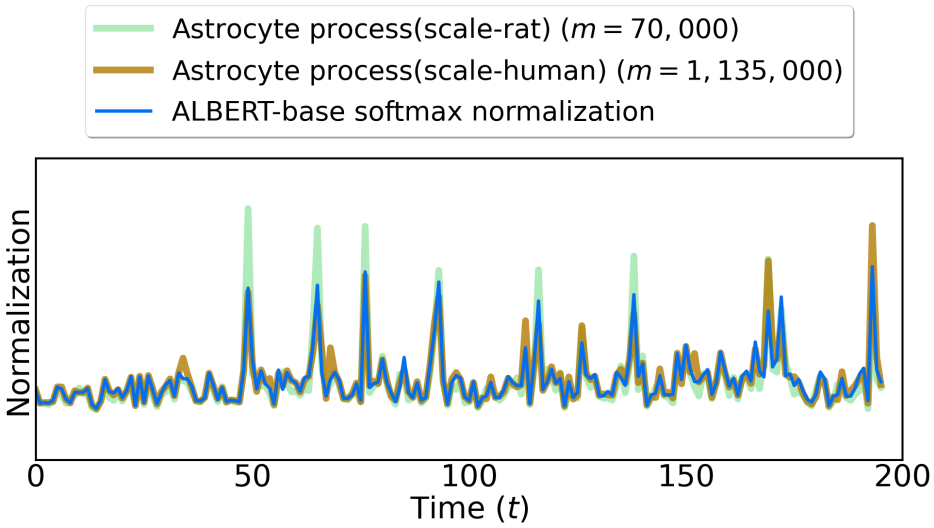

Figure 7: Using code from [1] this figure shows the normalization done by rodent astrocytes, human astrocytes and the softmax. It is clear that the human astrocyte process tends to align closer to the softmax but there are spikes in seemingly the right places for the rodent as well. This may not show anything at all but if one wanted to derive meaning, it could be argued that tuning the scale of the hidden neurons does not seem to correlate with the astrocyte scope. (self gen with their code)

5 Conclusion and Critical Analysis

The concept of a biological transformer architecture is fascinating! Given breakthroughs achieved by transformers recently, it makes sense to imagine a mechanism would exist in natural brains.

Starting with the main focus of this report: *Building transformers from neurons and astrocytes* by Kozachkov et al. [1], the biological and transformer background provided is a great introduction, but it requires some previous experience with transformers. The math is clearly presented, but there is no intuition which is given for the 'standard transformer' things. For example, if someone from a biological background is attempting to read without machine learning knowledge, the key, query, and value weights are probably confusing. Immediately they jump into self attention. The converse, biological primer for someone with a machine learning background seems adequate as they provide a nice list of points which are clear. These points are touched on throughout the paper and nicely identify why we need these features.

There are a lot of sources I reviewed, so I will just touch on some of the more significant ones.

Attention Is All You Need by google researchers [2]: it is clear why this paper is significant, they developed a novel method that significantly outperforms across the board. Funnily enough, I would not say it is a great introduction to transformers even though they invented the architecture. I would say the goal of this paper is to introduce the experts in the field to the transformer technology, but it is clear that they do not care to make the technology accessible. Some things were done such as Eq. [8] in their work which seems to be a somewhat crude justification for why $p_i\alpha$ goes to p but it seems to work in their simulation.

Visualizing transformers and attention by Grant Sanderson [8] is a stark contrast to the former paper. This is a talk given and posted to YouTube which concatenates the Neural Network video series by 3B1B (also Grant Sanderson) into a brief hour-long introduction to transformers. Each example is illustrative and when simplifications are made, they are for instructive purposes, and

the inaccuracies are noted. Reading *Attention Is All You Need*, after the intuitions provided by Sanderson, was a clearer experience. A small but important digression, I think recent societal aversion against education that has developed can be recognized in part due to the style of the former paper. Ensuring accessibility is exactly how one justifies to the layperson that investing in research is a worthy expenditure.

Much of the other work I reviewed were old biological papers about astrocyte function that I skimmed out of curiosity.

References

[1] L. Kozachkov, K. V. Kastanenka, and D. Krotov, "Building transformers from neurons and astrocytes," *Proceedings of the National Academy of Sciences*, vol. 120, no. 34, p. e2219150120, 2023.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), pp. 6000–6010, Curran Associates Inc., 2017.

[3] A. M. Smaldone, Y. Shee, G. W. Kyro, M. H. Farag, Z. Chandani, E. Kyoseva, and V. S. Batista, "A hybrid transformer architecture with a quantized self-attention mechanism applied to molecular generation," 2025.

[4] V. Porcu and A. Havlínová, "Past vs. present: Key differences between conventional machine learning and transformer architectures," *Advances in Nonlinear Variational Inequalities*, vol. 28, November 2024.

[5] E. A. Newman, "New roles for astrocytes: Regulation of synaptic transmission," *Trends in Neurosciences*, vol. 26, no. 10, pp. 536–542, 2003.

[6] G. Perea, M. Navarrete, and A. Araque, "Tripartite synapses: astrocytes process and control synaptic information," *Trends in Neurosciences*, vol. 32, pp. 421–431, Aug. 2009. Epub 2009 Jul 15.

[7] A. Araque, V. Parpura, R. P. Sanzgiri, and P. G. Haydon, "Tripartite synapses: glia, the unacknowledged partner," *Trends in Neurosciences*, vol. 22, no. 5, pp. 208–215, 1999.

[8] G. Sanderson, "Visualizing transformers and attention — talk for tng big tech day '24." `https://www.youtube.com/watch?v=KJtZARuO3JY`, 2024. Online Video, Accessed: 2025-06-03.

[9] N. A. Oberheim, T. Takano, X. Han, W. He, J. H. C. Lin, F. Wang, Q. Xu, J. D. Wyatt, W. Pilcher, J. G. Ojemann, B. R. Ransom, S. A. Goldman, and M. Nedergaard, "Uniquely hominid features of adult human astrocytes," *Journal of Neuroscience*, vol. 29, no. 10, pp. 3276–3287, 2009.

[10] L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, and R. B. Jackson, *Campbell Biology: A Global Approach*. Pearson, 12th ed., 2020. ISBN: 978-0135188743.

[11] M. M. Halassa, T. Fellin, H. Takano, J.-H. Dong, and P. G. Haydon, "Synaptic islands defined by the territory of a single astrocyte," *Journal of Neuroscience*, vol. 27, no. 24, pp. 6473–6477, 2007.

[12] G. R. J. Gordon, K. J. Iremonger, S. Kantevari, G. C. R. Ellis-Davies, B. A. MacVicar, and J. S. Bains, "Astrocyte-mediated distributed plasticity at hypothalamic glutamate synapses," *Neuron*, vol. 64, pp. 391–403, Nov 2009. Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't.

[13] F. Bullo, "Neural networks: Dynamics, stability, and optimization." Lecture notes, Department of Mechanical Engineering, UCSB, 2025. Version updated 2025/03/10.