Transformer as a hippocampal memory consolidation model based on NMDAR-inspired nonlinearity

Dong-Kyum Kim[1*]   Jea Kwon[2*]   Meeyoung Cha[1,3†]   C. Justin Lee[2†]

[1]IBS Data Science Group
[2]IBS Center for Cognition and Sociality
[3]KAIST School of Computing
{kdkyum,jeakwon,mcha,cjl}@ibs.re.kr

Abstract

The hippocampus plays a critical role in learning, memory, and spatial representation, processes that depend on the NMDA receptor (NMDAR). Inspired by recent findings that compare deep learning models to the hippocampus, we propose a new nonlinear activation function that mimics NMDAR dynamics. NMDAR-like nonlinearity shifts short-term working memory into long-term reference memory in transformers, thus enhancing a process that is similar to memory consolidation in the mammalian brain. We design a navigation task assessing these two memory functions and show that manipulating the activation function (i.e., mimicking the Mg^{2+}-gating of NMDAR) disrupts long-term memory processes. Our experiments suggest that place cell-like functions and reference memory reside in the feed-forward network layer of transformers and that nonlinearity drives these processes. We discuss the role of NMDAR-like nonlinearity in establishing this striking resemblance between transformer architecture and hippocampal spatial representation.

1 Introduction

Synaptic plasticity in the hippocampus is crucial for converting short-term memories into long-term memories during memory consolidation [1–3]. The N-methyl-D-aspartic acid receptor (NMDAR) is essential for mediating this memory formation as well as spatial representation. NMDAR serves as a switch for such plasticity and long-term memory formation [4–6]. Hippocampal place cells [7] and entorhinal cortex grid cells [8] are known to be central for spatial navigation in animals. NMDAR has been highlighted for its importance in place cell representations through hippocampal CA1 neurons [9, 3]. These discoveries have led to a deeper understanding of hippocampal function, inspiring recent efforts to replicate such spatial representation mechanisms in deep neural networks [10–12]. However, whether non-linear dynamics resembling hippocampal functions can be developed and used to support spatial representation in deep learning models remains unclear.

NMDAR is a post-synaptic ion channel that is characterized by nonlinear dynamics that distinguish it from other ion channels in the brain. These nonlinear dynamics are evident in the whole-cell current voltage (I-V) relationship (Fig. 1a) and are modulated by Mg^{2+} ion blockade at the pore region of the channel. Previous research indicates that the Mg^{2+}-dependent nonlinear dynamics of NMDAR play a key role in synaptic plasticity and memory formation [5, 6].

Recently, transformer-based deep learning models have been reported to have functions that resemble hippocampal formations [13]. Transformers comprise two consecutive modules: a self-attention
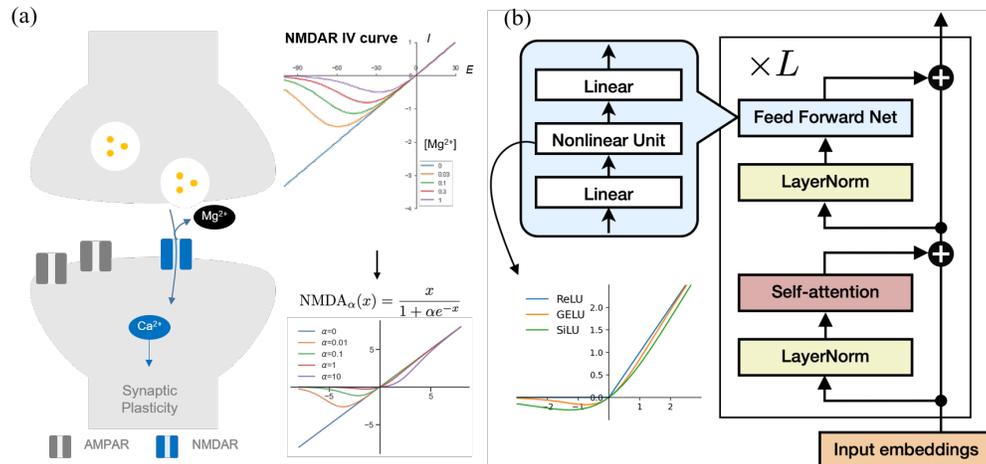
---

[*]Equal contribution.
[†]Corresponding authors.

Figure 1: (a) Schematic diagram of Mg^{2+}-gated NMDAR modulating synaptic plasticity (left), current-voltage dynamics (I-V curve; top right) and an NMDAR-inspired activation function, $\text{NMDA}_\alpha(x)$ (bottom right). (b) Transformer architecture and nonlinear activation functions in its feed-forward network (bottom left): ReLU, Gaussian Error Linear Unit (GELU), and Sigmoid Linear Unit (SiLU).

layer and a feed-forward network (FFN; see Fig. 1b). The self-attention layer is closely related to a recent neuroscience model that bridges with transformer [14] and it has been argued that softmax neurons in this layer behave like place cells in a navigation task [12]. However, the role of neurons in FFNs involved in a spatial navigation task has yet to be elucidated and compared to hippocampal characteristics.

This paper uncovers a resemblance between the NMDAR nonlinearity and recently developed activation functions commonly used in FFNs of deep learning models (Fig. 1). NMDAR functions operate on activity-dependent repulsion of Mg^{2+} ions [15, 16] and this phenomenon is particularly interesting because it supports self-gating of information flow (Ca^{2+} ion influx) in the post-synaptic region. Similar to this NMDAR activity-dependent gating mechanism, activation functions in modern neural networks combine input with a self-gating function (i.e., a sigmoidal function that ranges between 0 to 1) that determines information flow. Our observation prompted the following inquiry: **Can NMDAR-like nonlinearity in the feed-forward network layer of transformers enhance the formation of long-term memory and spatial representation by place cells?**

Here we derive a novel NMDAR-like activation function using the NMDAR current-voltage (I-V) curve. In parallel we design a spatial navigation task in a 2D grid environment that assesses two memory types well-defined by neuroscience research: working memory and reference memory [17, 18]. The former assesses events from within-trial, while the latter controls across trials in a stable (unchanging) environment. We then test the transformer model with the NMDAR-like activation function and find that 1) place cell representations emerge in the FFN layer rather than in the self-attention layer, 2) the nonlinearity of the NMDAR-like activation function can regulate reference memory, 3) place cell-like neurons in FFNs are strongly correlated with reference memory, whereas this correlation is not observed in the self-attention layer; and 4) NMDAR-like activation shows the best reference memory performance when compared to other available activation functions.

Collectively, these findings suggest that adopting NMDAR-like nonlinearity in FFN of transformer models can enhance the formation of long-term memory and spatial place cell representation, similar to previous observations in the animal brain. We believe these findings have important implications for developing new brain-inspired AI models and for understanding similar processes that occur in the brain and AI models.

2 Methods

2.1 Designing a 2D navigation task to test the role of working memory and reference memory

We designed a sensory observation prediction task in which an agent randomly walks in a 2D grid environment and is trained to predict subsequent sensory observations (see Fig. 2a) [12]. The agent
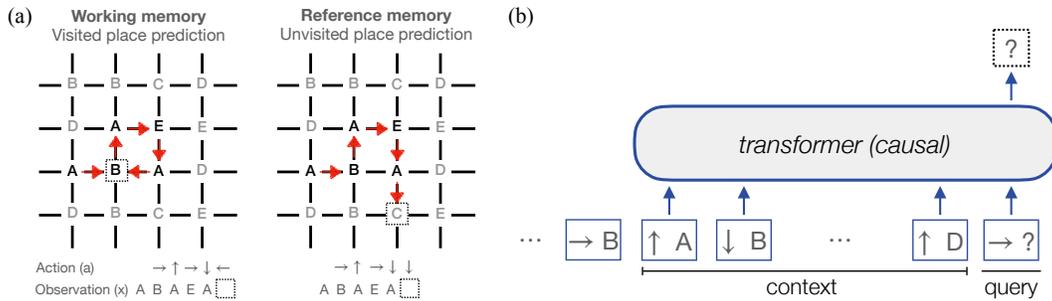
2

Figure 2: (a) Sensory observation prediction task in a 2D grid, where dotted squares indicate the target position to predict given a sequence of past actions and observations. The unvisited and visited places are represented in gray and black blocks, respectively. (b) A transformer model for predicting the location of an upcoming sensory observation based on sequences of [Action (a), Observation (x)] pairs. Using a sequence of pairs in the context, the model is trained to predict the masked observation (i.e., the subsequent observation) corresponding to the final query action.

receives a sequence of previous [Action (a), Observation (x)] pairs as input, and its goal is to predict the next observation, which is masked (i.e., dotted squares in the sequence of Observation (x) in the figure). We use the transformer architecture as our model.

We generated N maps of 11×11 2D grids. A random sensory observation chosen from ten letters is placed at each position on each map. Agents can move 'up,' 'right,' 'down,' 'left,' or 'stay.' An agent starts at a random position and initiates a random walk on the map, which is randomly selected from N training maps, for 2,048 steps for each trial.

Our design novelty is the consideration of two memory types: **short-term working memory** and **long-term reference memory**. When the prediction based on nodes that were previously visited during the random walk is incorrect, it counts as a *working memory error* (Fig. 2a left). In contrast, when the prediction based on unvisited nodes is incorrect, it counts as a *reference memory error* (Fig. 2a right). See details of the navigation task and definitions in Appendix A.4 and Fig. S3.

It is important to note that minimizing the reference memory error by memorizing input sequences is infeasible; the possible number of sequence configurations is exponential since the input sequence is randomly generated at each trial. To solve this task, the model needs to: 1) understand the abstract structure of 2D space, 2) infer which map it is on from the input sequence data, and 3) memorize the sensory observations and their positions on that map. Compared to the previous works [14, 12] that focused on working memory error, the current paradigm evaluates two distinct types of memory error and gives a more holistic view of model performance.

2.2 Transformer: two separate memory systems

We here review the self-attention mechanism and FFNs in the transformer architecture [13] that hypotheses about two separate memory systems are based on—working memory (formed within-trial) and reference memory (formed across-trials)—and why they are inferred to reside in self-attention layers and FFNs, respectively.

Self-attention mechanism   Given a sequence $\{\mathbf{x}_1, ..., \mathbf{x}_T\}$ of d-dimensional input embeddings, the self-attention layer calculates the interaction term between each embedding element within a context window via the self-attention mechanism. More formally, each input embedding applies two linear layers (W_k and W_v) to the embeddings to form the key matrix K and value matrix V:

$$K^\top = [\mathbf{k}_{t-c}^\top \; \mathbf{k}_{t-c+1}^\top \; \cdots \; \mathbf{k}_t^\top], \quad V^\top = [\mathbf{v}_{t-c}^\top \; \mathbf{v}_{t-c+1}^\top \; \cdots \; \mathbf{v}_t^\top], \tag{1}$$

where $\mathbf{v}_i = \mathbf{x}_i W_v$ ($W_v \in \mathbb{R}^{d \times d_k}$) and $\mathbf{k}_i = \mathbf{x}_i W_k$ ($W_k \in \mathbb{R}^{d \times d_k}$). Here, c denotes the context length. The key matrix $K \in \mathbb{R}^{(c+1) \times d_k}$ is then used to compute the interaction score between an input embedding at step t and all the vectors in K via dot products:

$$\mathbf{s}_t = \mathbf{q}_t K^\top, \quad \text{where } \mathbf{q}_t = \mathbf{x}_t W_q \; (W_q \in \mathbb{R}^{d \times d_k}). \tag{2}$$

The normalized values of $\mathbf{s}_t \in \mathbb{R}^{(c+1)}$, called attention values, are calculated via the softmax function; the final output of the self-attention mechanism is a weighted sum of the value vectors in

Table 1: Comparison of common activation functions with $\text{NMDA}_{\alpha,\beta}$

$\text{NMDA}_{\alpha,\beta}$	Name	Equation	Reference
$\text{NMDA}_{\alpha=1,\beta=1}(x)$	SiLU(x)	$x\sigma(x)$	[20]
$\text{NMDA}_{\alpha=1,\beta=1.702}(x)$	GELU(x)	$x\sigma(1.702x)$	[21]
$\text{NMDA}_{\alpha=1,\beta=\infty}(x)$	ReLU(x)	$\max(0,x)$	[22]
$\text{NMDA}_{\alpha=1,\beta}(x)$	Swish(x)	$x\sigma(\beta x)$	[23]

$V \in \mathbb{R}^{(c+1)\times d_k}$ with the attention values:

$$\mathbf{y}_t = \texttt{softmax}\left(\frac{\mathbf{q}_t K^\top}{\sqrt{d_k}}\right) V. \tag{3}$$

After this update, $\mathbf{y}_t \in \mathbb{R}^{d_k}$ is updated by another linear transformation $W_o \in \mathbb{R}^{d_k \times d}$: $\mathbf{z}_t = \mathbf{y}_t W_o$. The output $\mathbf{z}_t$ is added to the $\mathbf{x}_t$; $\mathbf{z}_t + \mathbf{x}_t$ providing the final output of the self-attention layer, and this information is sent through to the subsequent layer.

Feed-forward networks (FFNs) This component consists of two linear layers with a point-wise nonlinear activation function ϕ:

$$\texttt{FFN}(\mathbf{x}_t) = \phi(\mathbf{x}_t U_1^\top)U_2, \tag{4}$$

where $U_1 \in \mathbb{R}^{d_f \times d}$ and $U_2 \in \mathbb{R}^{d_f \times d}$ are trainable weight matrices. Sukhbaatar et al. [19] showed that Eq. (3) and Eq. (4) have similar structures except for the following: 1) U_1 and U_2 matrices are fixed over different input sequences while K and V matrices are dynamically changed with input and 2) operations in FFNs are point-wise or local while the self-attention layer has non-local operations, e.g., the softmax function and dot products between different elements. This observation suggests that the FFNs store "general" knowledge about the task that does not depend on the situation.

2.3 Resemblance of NMDA receptor nonlinear dynamics with modern activation functions

NMDAR's nonlinear dynamics mostly arise from the voltage-gated Mg^{2+} repulsion at the NMDAR channel's pore [15, 16] (Fig. 1a left). Previous work showed this nonlinear I-V relationship to be:

$$I_{\text{norm}} = V\mathbf{p}_{\alpha,\beta}(V) \tag{5}$$

where V represents an input voltage and $\mathbf{p}_{\alpha,\beta}(V)$ is a voltage-dependent channel open probability that follows the ion blockade model [24]:

$$\mathbf{p}_{\alpha,\beta}(V) = \frac{1}{1 + \alpha e^{-\beta V}} \tag{6}$$

where $\alpha = [Mg^{2+}]/K_{Mg^{2+}}$ is a parameter determined by $[Mg^{2+}]$, $K_{Mg^{2+}}$ is a dissociation constant, and β is a temperature constant. For further details, see Appendix A.2.

The dynamics of the NMDA receptor closely resemble those of new activation functions such as ReLU or GELU (see Fig. S2 in Appendix A.3). This visual resemblance motivated us to define a new NMDAR-inspired activation function (see details in Appendix A.3) as follows:

$$\text{NMDA}_{\alpha,\beta}(x) = x\mathbf{p}_{\alpha,\beta}(x) = \frac{x}{1 + \alpha e^{-\beta x}}. \tag{7}$$

This $\text{NMDA}_{\alpha,\beta}(x)$ incorporates modern activation functions with varying values of temperature constant, β (Table 1). To investigate the Mg^{2+}-gated nonlinear dynamics, α, we compared various activation functions with $\text{NMDA}_\alpha(x) = x\mathbf{p}_{\alpha,\beta=1}(x)$.

2.4 Testing the NMDAR-inspired activation in navigation tasks

The transformer model is trained using softmax cross-entropy loss to predict subsequent sensory observations (i.e., dotted squares in Fig. 2). Instead of using sinusoidal positional encoding [13], we employ recurrent positional embedding which encodes the location of an input element by using the recurrent neural network (RNN); this method is closely related to the most advanced neuroscience model of the hippocampus [12].

We generate the embedding vectors of the sensory observation sequence with a word embedding layer, but the embedding vectors of the action sequence are generated using an RNN; $\mathbf{e}_{t+1} = \tanh(\mathbf{e}_t W_a)$, where $\mathbf{e}_t$ is a recurrent positional embedding at step t, and W_a is the action-dependent trainable weight matrix. The input is given by $\{[\mathbf{e}_1, \mathbf{x}_1], [\mathbf{e}_2, \mathbf{x}_2], \ldots, [\mathbf{e}_t, \mathbf{x}_t]\}$, where $\mathbf{x}$ denotes the embedding vector of sensory observation x; the initial recurrent positional embedding $\mathbf{e}_1$ is sampled from a normal distribution, and we mask the last observation x_t.

In our experiment, FFN in the transformer model consists of two linear layers (see Fig. 1b and Eq. (4)) with the NMDAR-inspired activation function NMDA_α. We use TransformerXL [25] with an extended memory length of 32 and segment length of 32 so that the context length c is 64 and working memory error is measured when the node to predict its sensory observation is in the context window (see Fig. 2b); i.e., a node that the agent has not been visited in the last 64 steps is treated as *an unvisited node*. Note that our model is unable to access the sensory observations of unvisited nodes via the self-attention mechanism due to the fixed context window size.

The input embedding is a concatenated vector $[\mathbf{e}, \mathbf{x}]$ of the word embedding $\mathbf{x}$ (dimension of 256), and the recurrent positional embedding $\mathbf{e}$ (dimension of 256) so that the total input embedding dimension is 512. The number of heads in the self-attention layer is 8, and the number of neurons in the FFN is 2,048. The dropout rate is set to 0.1, and the maximum clip norm of the gradient is set to 0.25. We employ the ADAM optimizer [26] and a learning rate schedule with a linear decay from 0.0001 (start) to 0 (end). We run 512 random walk simulations (trials) in parallel to collect training trajectories. The total number of random walking steps is 2,048 for each simulation, and the total number of steps for training a model is 512 (batch size; the number of trials per epoch) $\times$ 2,048 (total number of steps in a trial) $\times$ 200 (number of epochs) (see Fig. S3 in Appendix A.4). The average number of unvisited nodes in a single trial was 561. The PyTorch code for reproducing our results is available at `https://github.com/kdkyum/transformer_memory_consolidation`.

3 Results

3.1 Working memory and reference memory errors

To measure the impact of nonlinearity α in the FNNs, we trained transformer models with different values of α in $[0, 0.01, 0.05, 0.1, 0.5, 1, 5, 10]$ and evaluated the working memory and reference memory errors on the train maps (i.e., familiar maps) and test maps (i.e., novel maps). The changing α values in a transformer model mimic the changes in NMDAR Mg^{2+}-gating in the brain, as inspired by neuroscience findings that selective inhibition of hippocampal CA1 NMDAR can disrupt the consolidation of short-term working memory to long-term reference memory.

The top left plot in Fig. 3a shows that the reference memory error on the training maps rapidly decreased over training trials when α is larger than zero, with a larger improvement for increasing α. The reference memory error on the novel maps, however, was nearly constant at a chance level of 0.9 ($= 1 - 1/(\text{number of letters})$) for all α (see Fig. 3a top right). Fig. 3a bottom right shows that working memory was active on novel maps that had not previously been shown during training. This finding suggests that working memory formation is intact on novel maps. Training the models on different numbers of maps, N, Fig. 3b shows that increasing nonlinearity (i.e., α) helped to activate reference memory, and the trend of improvement was consistent, as shown for $N = 32$, 48, and 64. Training more maps led to larger reference memory errors, because increasing N means the model needs to memorize what-where pairs (i.e., each training contains unique what-where information).

In addition, we find that removing nonlinearity in the NMDAR-inspired activation function ($\alpha = 0$) impairs reference memory formation (Fig. 3b) but leaves working memory formation intact (Fig. 3a). This means that even though the trainable parameters exist in the self-attention layer, a lack of nonlinearity in the feed-forward layer significantly impairs the formation of long-term reference memory in familiar maps. This result suggests that short-term working memory and long-term reference memory are physically stored in separate structures (the self-attention layer and the FFN) of transformers and possibly gated by the nonlinear activation functions residing in the FFN.

We next tested other nonlinear activation functions, including GELU ($x\sigma(1.702x)$), ReLU ($\max(0, x)$), LeakyReLU ($\max(0, x) + 0.01 \min(0, x)$), sigmoid, and tanh. Fig. 3c shows that the newly proposed biologically-inspired function ($\text{NMDA}_{\alpha=10}$) yielded substantially lower reference memory errors relative to other activation functions. A comparison with the second-best-performing

GELU activation, for example, indicated a statistically significant difference with p-value less than 0.05 (see Appendix A.6). These results open up the possibility of finding a better activation function through a new α hyperparameter space of the NMDA-inspired activation function.

Expanding on the evaluation, one may consider other memory types like recurrent positional embedding instead of reference memory. We used non-recurrent positional embeddings to train the models (see Appendix A.5) and confirmed that working memory and reference memory errors increased more substantially. This finding supports the idea that working memory is crucial for memory consolidation and that disrupting it impairs reference memory. However, we also saw a similar trend of decreasing reference memory error while increasing α of NMDA_α (see Fig. S4).

We further assessed the prediction error of the first visited node. While the reference memory error is defined as a prediction error on a node that the agent has not visited in the previous 65 steps, the first visited node prediction error is a prediction error on a node that the agent visits for the first time in a trial. The results for the first visited node prediction error were identical to the results for the reference memory error (see Fig. S6 in Appendix A.5). These findings suggest that reference memory is used in training maps to predict the unvisited node.

3.2 Place cells in feed-forward networks

One of the most striking characteristics of the hippocampus is the presence of the **place cells** [7], which are neurons in the brain that fire at a particular place in the environment [7]. Studies have shown that hippocampal place cells encode spatial location through localized firing patterns. They have been considered a substrate for long-term memory of the location where specific events occurred (e.g., a previously visited position in a navigation task). Selective impairment of NMDAR in hippocampal CA1 disrupts place cell emergence and memory consolidation of long-term memory [2, 3, 9].

We investigated whether place cells emerge during the formation of long-term reference memory in a transformer model, building upon our previous results demonstrating selective impairment in long-term reference memory with NMDA_α modulation. We newly designed a metric called the **place cell score** described in Eq. S1 in Appendix A.1 and examined the spatial representation within the transformer architecture. This metric, ranging from 0 (i.e., homogeneous firing) to 1 (i.e., firing is specific to the agent's position in the grid) quantifies the firing specificity of each neuron for spatial
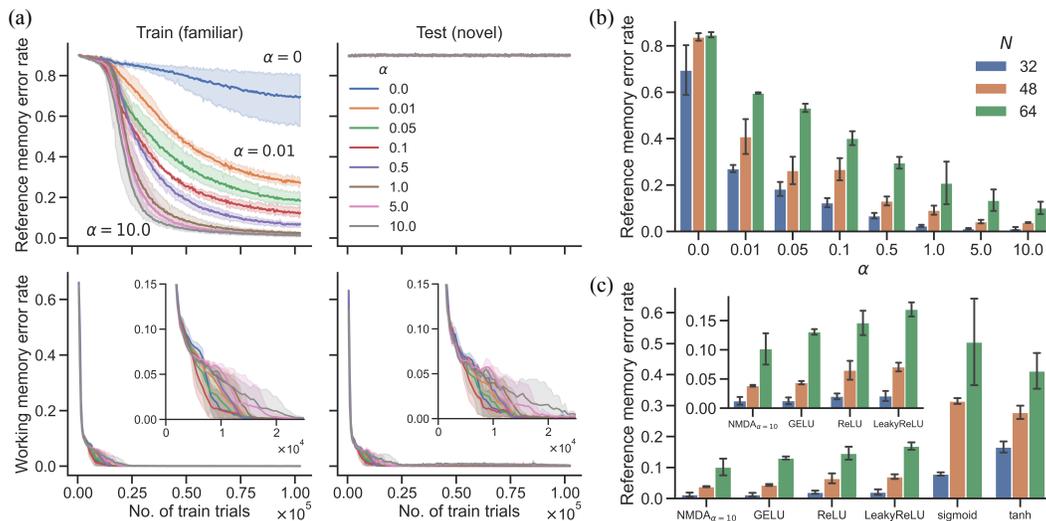

Figure 3: (a) Reference and working memory error rates over training trials for training maps and testing maps for $N = 32$ training maps. Insets in the bottom figures show working memory error rates during the initial training phase. (b) Reference memory errors were evaluated on training maps over different values of α in NMDA_α and N. (c) Reference memory error comparison between $\text{NMDA}_{\alpha=10}$, GELU, ReLU, LeakyReLU, sigmoid, and tanh activation functions. Inset: magnified view of the top 4 activation functions. Error bars and shaded areas represent the standard deviation of errors from three independently trained models.
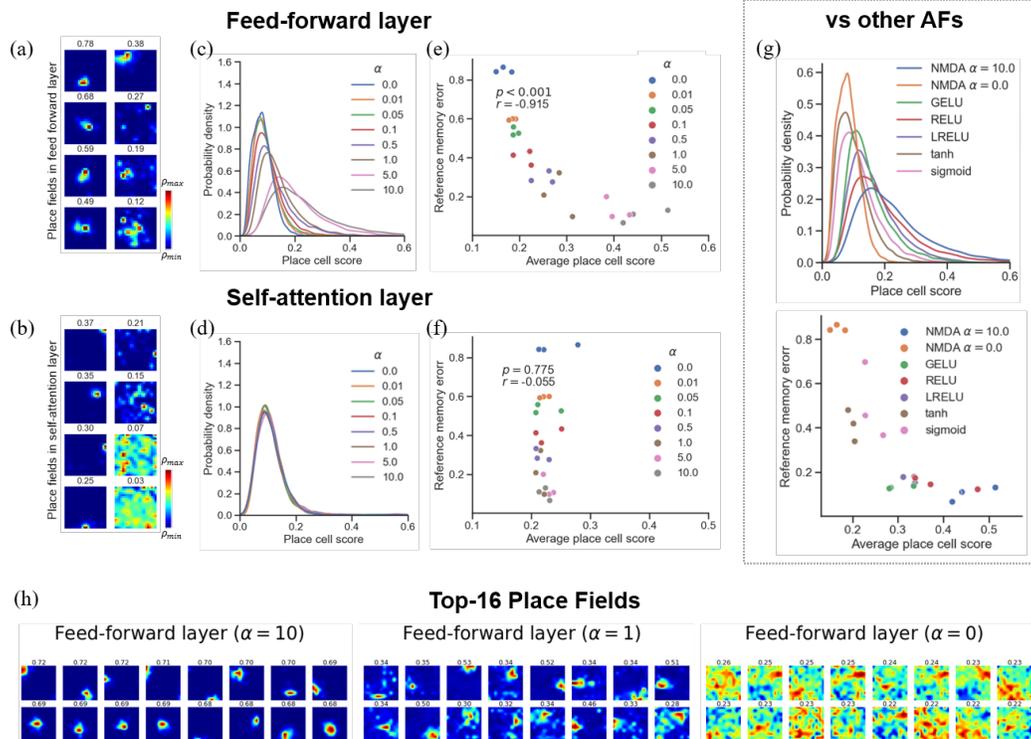
Figure 4: Reference memory-related place cells selectively emerge in FFN but not in the self-attention layer over increasing α. (a, b) Example rate maps with place scores in FFN and self-attention layers at $\alpha = 10$; from top left (high) to bottom right (low); color bar indicates the firing rate between ρ_{max} and ρ_{min}. (c-d) Place cell score distributions with varying α in FFN (c) and self-attention layers (d). (e-f) Scatter plot of average place cell scores and reference memory errors. r and p denote Spearman's rank correlation coefficient and significance score, respectively. (g) place cell score distribution and relationship of average place cell scores and reference memory errors in common activation functions: GELU, ReLU, LeakyReLU, tanh, and sigmoid. All results are evaluated using training maps. (h) Rate maps of neurons with top-16 place cell scores in the FFNs with varying values of α; $\alpha = 10$ (left), $\alpha = 1$ (middle), and $\alpha = 0$ (right).

locations, similar to the approach described in the TEM-t model [12]. We recorded the activation values of each neuron at every step during a random walk process. As the agent traversed the environment, which had an 11×11 grid structure, we accumulated the activation values for each grid point, resulting in a 2D array known as the *rate map* or *place field* (i.e., the spatial distribution of single neuron activation during a random walk, as shown in Fig.4a and Fig.4b).

Fig. 4a and 4b show the rate maps of neurons with place cell scores in the FFN and self-attention layers, respectively. For the self-attention layer, the total number of neurons in the softmax layer was 65 (context length + masked sensory observation) $\times$ 8 (number of heads) $\times$ 2 (number of layers). The total number of neurons in the FFN layer was set at 2,048 (number of neurons) $\times$ 2 (number of layers). As can be seen, our metric accurately represents location specificity. Fig. 4c and 4d show the distribution of place cell scores in the two layers with different values of α. When the α value is increased, the place cell score distribution found in the FFN becomes positively shifted (see Fig. 4h rate map examples for $\alpha = 0, 1.0$, and 10.0), whereas the place cell score distribution in the self-attention layers remains.

Fig. 4e and 4f show a relationship between the average place cell score and the reference memory error for each α. The average place cell scores in the self-attention layers show no correlation with reference memory errors, while neurons in the FFN exhibit a clear correlation. These results imply that reference memory formation and place cell emergence can be enhanced by NMDAR-like nonlinearity in the FFN.

Fig. 4g compares the place cell representations of our NMDA ($\alpha = 0, 10$) with the representations in FFNs with the activation functions used in Fig. 3c, indicating that the case of NMDA$_{\alpha=10}$ outperforms other activation functions, in both reference memory formation and place cell representation. Our finding that large α (the [Mg^{2+}] component) enhances reference memory is in line with the biological observation that increasing the [Mg^{2+}] in the brain enhances long-term memory formation [6].

We also explored the possibility of extending the hyper-parameter space by contrasting the NMDA-inspired activation function with an alternative activation function based on α. LeakyReLU was considered for this experiment because of its adjustable negative slope, which was re-defined as $(\max(0, x) + \alpha \min(0, x))$. The results reported in Fig. S10 of Appendix A.7 show that LeakyReLU activation, for all tested α values, produced lower place cell scores than the best-performing NMDA-inspired function (i.e., NMDA$_{\alpha=10}$). Our findings suggest that NMDA-like nonlinearity, which is already known to be critical in biological systems, could be used to enhance transformer models.

Building upon these results, our work connects memory consolidation theory from neuroscience to transformer-based processing of short-term and long-term memory. Unlike previous studies that directly provide 2D geometric information to models [11, 10], our model only takes sensory-action pairs as an input, which was sufficient to observe the emergence of place cells in the transformer model. Specifically, we showed that nonlinearity in the transformer's FFNs plays a crucial role in transforming short-term working memory from the self-attention layer into long-term memory, and the place cell emergence. The observed place cells may correspond to sparse interpretable activation in FFN. See Appendix A.10 for further analysis on the sparsity and place cells. These findings provide new insight into the inner workings of transformers and their connection to neural processes in the brain.

4 Related works

The current study is inspired by recent studies that connect neuroscience and AI models. One such seminal work is by Whittington et al. [12], which showed a relationship between the self-attention layer and a recent hippocampal model called the Tolman-Eichenbaum Machine (TEM; Whittington et al. [14]). Our work expands the literature by focusing on FFNs in the transformer and making a connection to the emergence of place cells.

TEM is a neuroscience-based model that reproduces neural representations in the hippocampus and entorhinal cortex. Instead of storing memory in the key matrix K and value matrix V, it instead stores memory using a Hebbian weight matrix $M \in \mathbb{R}^{d_k \times d_k}$. Every outer product of key and value vector $\mathbf{k}_i^\top \mathbf{v}_i$ at each step i is simply stored in M via the Hebbian update rule. M is initialized to a zero matrix at the beginning of the task and adds every outer product at each time step:

$$M = a \sum_{i=1}^{t} \mathbf{k}_i^\top \mathbf{v}_i = a K^\top V, \tag{8}$$

where a is a weighting factor. In the memory retrieving phase with the query vector $\mathbf{q}$, TEM uses an attractor network:

$$\mathbf{q}M = a\mathbf{q}K^\top V. \tag{9}$$

Whittington et al. [12] found that the memory retrieving process in TEM has a close mathematical structure to Eq. (3) when the softmax function is replaced with a linear function. We note that methods for linearizing softmax in self-attention layers have been studied to address high computational costs when context length c is very long [27, 28]. The subsequent model of TEM, called TEM-t [12], replaces the attractor network (Eq. (9)) with a self-attention mechanism (Eq. (3)). This study demonstrated that TEM-t learns significantly faster than TEM.

TEM-t and TEM do not have a fixed context length c; therefore, these models store all information before step t, i.e., $c = t$. The computational cost of the self-attention layer in TEM-t is $O(t^2)$, and retaining all previous information is too expensive from both a biological and computational standpoint[3]. For TEM, the Hebbian update rule has no quadratic computational cost and can add all previous information in a fixed number of synapses d_k^2; however, the memory capacity of the Hebbian matrix M is $O(d_k)$ and the speed of memory retrieval is substantially slower than the

---

[3]Due to the computational cost, TEM-t does not store all historical data but selectively chooses which data to store in K and V.

self-attention mechanism [29–31]. In contrast to TEM and TEM-t that rely on a single memory system, we investigated two separate memory systems: 1) context-dependent matrices K and V in the self-attention layer with a fixed context length c and 2) context-independent fixed matrices U_1 and U_2 (in Eq. (4)) in the FFNs.

Our research differs from previous studies in the following ways: 1) We designed a navigation test that assesses working memory and reference memory separately, providing a more comprehensive evaluation of the model's performance. 2) We proposed a new brain-inspired activation function, NMDA_α, which relates to modern nonlinear activation functions and allows for the analysis of the effect of α on long-term reference memory formation. 3) We demonstrated that place cell-like neurons emerge in FFNs in conjunction with reference memory formation, which is a novel finding that has not been addressed in the TEM or TEM-t models. 4) TEM and TEM-t focus on working memory errors and do not cover reference memory errors. In contrast, our work evaluates both types of memory errors, providing a more detailed analysis of the model's performance.

5 Discussion

While extensive efforts have been directed toward finding the optimal nonlinear activation function for improving modern deep neural network models [21–23], their relationship to neural substrates that mediate nonlinearity in the human brain remains obscure. Furthermore, the role of nonlinearity in intelligent functions remain unclear. Our research attempts to fill this gap by proposing and testing how a biologically inspired nonlinearity functions in a transformer model that was previously related to the hippocampal formation. We examined functions in terms of long-term reference memory formation and place cell representation. This idea was evaluated in a carefully designed 2D grid environment and by implementing an activation function derived from NMDAR-like nonlinearity.

Our research reveals that place cell-like neurons that are critical to spatial navigation can be found in both the self-attention layers and FFNs of the transformer model, similar to the presence of place cells in the CA3 and CA1 regions of the hippocampus [32]. In the hippocampus, the CA3 region is thought to be involved in the initial formation of new memories, specifically in pattern completion [33], while CA1 is thought to be important for the long-term consolidation of memories. As such, we suggest that the CA3 region may serve a similar function to the self-attention layer, while the CA1 region may function similarly to FFN. However, further research is needed to fully understand the similarities and differences between the properties of place cells in the transformer model and those found in the CA3 and CA1 regions of the hippocampus. For more discussions regarding the biological plausibility of this proposal, see Appendix A.8.

Recent machine learning research has tested whether the transformer architecture is analogous to different types of biological memory. It has been suggested that (1) transformer FFN modules resemble associative memory [34], (2) the FFN in a transformer block functioning as a key-value memory [35, 36], (3) activation sparsity in the transformer FFN enhances robustness to noisy input [37, 38], and (4) sparse activity in FFNs increases the percentage of neurons that selectively activate to human interpretable input features [39].

Our data instead suggests the transformer architecture resembles *memory consolidation* by the animal brain, which refers to the transfer process of a short-term memory into a long-term memory system in neuroscience research [40]. Previous research has revealed that Mg^{2+}-gating of NMDA receptors modulates the formation of long-term memories [5, 6]. These observations imply that the nonlinear dynamics of NMDA receptors in hippocampus CA1 are critical for consolidating short-term memory into long-term memory.

To our surprise, our results agree qualitatively with previous NMDAR impairment experiments from neuroscience: 1) selective inhibition of hippocampal CA1 NMDAR inhibition does not disrupt working memory [41] but impairs the long-term memory formation [2], 2) changing NMDAR Mg^{2+}-gating (changing α in this work) enhances or disrupts long-term memory formation [5, 6], 3) NMDAR is required for long-term stabilization of newly forming place fields [9, 3]. These similarities between hippocampal memory consolidation and our results suggest that the transformer is an effective memory consolidation model.

Our research points to exciting future directions. The current study examined what-where memory using a sensory observation task in a static environment. However, our real-world environment

changes constantly and provides new inputs over time. Modern deep learning systems are generally incapable of adapting to a dynamic environment or reordering of sensory inputs. We intend to explore what-where-when memory, called *episodic memory*, in transformers and other deep models.

Acknowledgments and Disclosure of Funding

References

[1] Fei Li and Joe Z Tsien. Memory and the nmda receptors. *The New England journal of medicine*, 361(3):302, 2009.

[2] Joe Z Tsien, Patricio T Huerta, and Susumu Tonegawa. The essential role of hippocampal ca1 nmda receptor–dependent synaptic plasticity in spatial memory. *Cell*, 87(7):1327–1338, 1996.

[3] Clifford Kentros, Eric Hargreaves, Robert D Hawkins, Eric R Kandel, Matthew Shapiro, and Robert V Muller. Abolition of long-term stability of new hippocampal place cell maps by NMDA receptor blockade. *Science*, 280(5372):2121–2126, 1998.

[4] Tim VP Bliss and Graham L Collingridge. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361(6407):31–39, 1993.

[5] Inna Slutsky, Nashat Abumaria, Long-Jun Wu, Chao Huang, Ling Zhang, Bo Li, Xiang Zhao, Arvind Govindarajan, Ming-Gao Zhao, Min Zhuo, et al. Enhancement of learning and memory by elevating brain magnesium. *Neuron*, 65(2):165–177, 2010.

[6] Tomoyuki Miyashita, Yoshiaki Oda, Junjiro Horiuchi, Jerry CP Yin, Takako Morimoto, and Minoru Saitoe. Mg2+ block of drosophila NMDA receptors is required for long-term memory formation and CREB-dependent gene expression. *Neuron*, 74(5):887–898, 2012.

[7] John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971.

[8] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.

[9] Thomas J McHugh, Kenneth I Blum, Joe Z Tsien, Susumu Tonegawa, and Matthew A Wilson. Impaired hippocampal representation of space in ca1-specific nmdar1 knockout mice. *Cell*, 87 (7):1339–1349, 1996.

[10] Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

[11] Christopher J. Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B17JTOe0-.

[12] James C. R. Whittington, Joseph Warren, and Tim E.J. Behrens. Relating transformers to models and neural representations of the hippocampal formation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=B8DVo9B1YE0.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[14] James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

[15] L Nowak, P Bregestovski, P Ascher, A Herbet, and A Prochiantz. Magnesium gates glutamate-activated channels in mouse central neurones. *Nature*, 307(5950):462–465, 1984.

[16] Mark L Mayer, Gary L Westbrook, and Peter B Guthrie. Voltage-dependent block by mg2+ of nmda responses in spinal cord neurones. *Nature*, 309(5965):261–263, 1984.

[17] David S Olton, Christine Collison, and Mary Ann Werz. Spatial memory and radial arm maze performance of rats. *Learning and motivation*, 8(3):289–314, 1977.

[18] David S Olton, James T Becker, and Gail E Handelmann. Hippocampus, space, and memory. *Behavioral and Brain sciences*, 2(3):313–322, 1979.

[19] Sainbayar Sukhbaatar, Edouard Grave, Guillaume Lample, Herve Jegou, and Armand Joulin. Augmenting self-attention with persistent memory, 2019. URL https://arxiv.org/abs/1907.01470.

[20] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

[21] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[22] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[23] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[24] Ann M Woodhull. Ionic blockage of sodium channels in nerve. *The Journal of general physiology*, 61(6):687–708, 1973.

[25] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Association for Computational Linguistics*, pages 2978–2988, 2019.

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6980.

[27] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

[28] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9355–9366. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/schlag21a.html.

[29] Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, may 2017.

[30] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

[31] Dmitry Krotov and John J. Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=X4y_10OX-hX`.

[32] Can Dong, Antoine D Madar, and Mark EJ Sheffield. Distinct place cell dynamics in ca1 and ca3 encode experience in new environments. *Nature communications*, 12(1):2977, 2021.

[33] Kazu Nakazawa, Michael C Quirk, Raymond A Chitwood, Masahiko Watanabe, Mark F Yeckel, Linus D Sun, Akira Kato, Candice A Carr, Daniel Johnston, Matthew A Wilson, et al. Requirement for hippocampal ca3 nmda receptors in associative memory recall. *science*, 297 (5579):211–218, 2002.

[34] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022.

[35] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[36] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

[37] Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J Reddi, Ke Ye, Felix Chern, Felix Yu, Ruiqi Guo, et al. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. In *The Eleventh International Conference on Learning Representations*.

[38] Sebastian Jaszczur, Aakanksha Chowdhery, Afroz Mohiuddin, Lukasz Kaiser, Wojciech Gajewski, Henryk Michalewski, and Jonni Kanerva. Sparse is enough in scaling transformers. *Advances in Neural Information Processing Systems*, 34:9895–9907, 2021.

[39] Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer ElShowk, Nicholas Joseph, Nova DasSarma, Ben Mann, Danny Hernandez, Amanda Askell, Kamal Ndousse, Andy Jones, Dawn Drain, Anna Chen, Yuntao Bai, Deep Ganguli, Liane Lovitt, Zac Hatfield-Dodds, Jackson Kernion, Tom Conerly, Shauna Kravec, Stanislav Fort, Saurav Kadavath, Josh Jacobson, Eli Tran-Johnson, Jared Kaplan, Jack Clark, Tom Brown, Sam McCandlish, Dario Amodei, and Christopher Olah. Softmax linear units. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/solu/index.html.

[40] James L McGaugh. Memory–a century of consolidation. *Science*, 287(5451):248–251, 2000.

[41] Inah Lee and Raymond P Kesner. Differential contribution of nmda receptors in hippocampal subregions to spatial working memory. *Nature neuroscience*, 5(2):162–168, 2002.

A   Appendix

A.1   Place cell

A.1.1   Place cell overview

Place cells are a distinct class of neurons in the hippocampus that fire selectively in a certain location [S1], and provide a sparse population code for self-location. Place cells play a key role in understanding how the hippocampus represents spatial locations [S2], which is characterized by their place fields, which are spatially stable regions where the cell preferentially fires. As the subject moves through the environment, different place cells fire in response to its position, collectively forming a cognitive map [S3]. This dynamic encoding allows the brain to create and maintain spatial representations, which are essential for navigation and memory [S4].

Relevance to machine learning   Accurately identifying place cells and understanding their properties is crucial for machine learning practitioners working on spatial representation, navigation, and memory problems. Recent studies investigated the role of place cells and grid cells in spatial representation in navigation tasks with deep models [S5–S8]. The concept of place cells can provide insights that support the development of novel algorithms and techniques for potential machine learning applications such as robotics, autonomous vehicles, virtual reality, and other areas where spatial understanding is essential.

Grid cells and path integration   While our work was mainly on the emergence of place cells, we could also observe grid cells based on the experiment, similar to TEM and TEM-t models. When we removed the recurrent connection in the action embedding layer, grid cells did not emerge, and the model's performance declined, indicating the role of grid cells in place cell formation and spatial memory. An interesting hypothesis in neuroscience is that the interaction between grid cells, path integration, and place cells forms a critical loop [S9, S10] in the neural circuitry underlying spatial memory formation, which is similar to our findings from removing recurrent connections. This finding awaits future investigation on how grid cells and path integration affect place cell emergence and spatial memory formation.
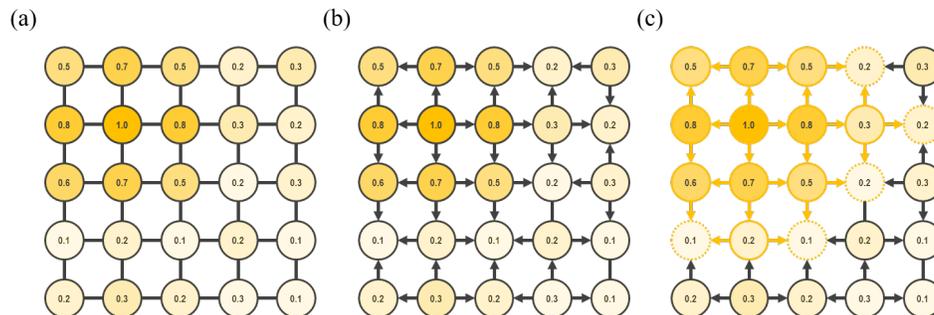
A.1.2   Place cell score calculation

Figure S1: Schematic diagram for creating auxiliary graph for place cell score calculation (a) Consider the 2D grid environment of the place field as an undirected graph $\mathcal{G}$. Colors represent the firing rate of nodes. (b) Convert $\mathcal{G}$ into a directed graph by comparing the firing rate ρ between nodes (direction of higher to lower firing rates) (c) Operate Depth First Search algorithm from peak node to construct all connected components }(yellow). Dashed circles are leaf nodes.

Our place cell score metric is inspired by the metric described in the TEM-t model [S8]. The place cell score in Eq. S1 is a scalar value that quantifies each neuron's firing specificity for a grid location. We record every neuron's activation value at every step during the random walk process. Since the map has an 11×11 grid structure, we accumulate the activation value for each grid point as the agent random walks, creating a 2D array. This array is called the rate map or the place field (i.e., the spatial distribution of single neuron activation during a random walk).

We investigated the role of neurons in the FFNs and self-attention layers by measuring their place specificity. Given a $K \times K$ 2D grid environment as graph $G = (V, E)$ and a firing rate (i.e., cumulative activation value at node i divided by the length of evaluation trial) of node $i \in V$ as a ρ_i, we defined a *maximally firing node* as $i_{\max}$ and its firing rate as $\rho_{\max}$. E are directed edges that connect high to low firing nodes in G. From G, we ran a depth-first-search from source node $i_{\max}$ to build a sub-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ which we call *all connected components* S1. Given G and $\mathcal{G}$, the *place cell score* is defined as follows:

$$\text{Place cell score} = \gamma \frac{\sum_{i \in \mathcal{V}} \rho_i}{\sum_{i \in V} \rho_i}, \tag{S1}$$

where $\gamma = 1 - |\mathcal{V}^*|/|V|$ is a discount factor and $\mathcal{V}^*$ is $\mathcal{V}$ without node $i_{\max}$ and leaf nodes. To measure place cell score, we recorded the firing rate ρ_i of neurons over a random walking trajectory with 10^5 steps in one of the training maps. We then measured the place cell scores of neurons in the FFN and self-attention layers. The place cell score is 1 when the neuron is firing only at a specific node; the score is 0 when it fires homogeneously across all nodes.

Since 2D-grid spaces cannot be directly interpreted as Euclidean spaces, it is not appropriate to apply space-related functions, such as Gaussian kernel density estimation, to the place fields. Place cell evaluation methods in previous research can be broadly categorized into four types [S11]:

1. **Peak method:** Classifies cells based on the average firing rate in one location being higher than in the rest of the environment.

2. **Stability method:** Classifies cells with stable firing patterns across locations over time.

3. **Information method:** Classifies cells based on the increased amount of spatial information they hold about the animal's location.

4. **Combination method:** A combined approach that considers cell's place fields, including their size, peak, and activity.

We chose the Peak method approach because it easily applies to 2D-grid structures.

To implement the Peak method in a 2D grid environment, we represent the 2D grid environment as a graph structure where neuron activities at each node are tracked while the agent explores the environment. Averaging neuronal activities per node yields the neuron's place fields, as illustrated in Figures 4 and 5. In these place fields, we find a node with a maximal firing rate, $i_{\max}$ Starting from this node, we find all connected components as a subgraph by connecting neighboring nodes within the nearest minimum. We create directed edges connecting nodes from high to low firing rates and run depth first search (DFS) from the node $i_{\max}$. DFS can be considered an inverted watershed algorithm that fills a single pool containing the global minimum. The advantage of our place cell metric is its ability to generalize to various graph structures, such as 2D hexagonal grids or 3D grid spaces.

A.1.3 Pseudo code for calculating place cell score metric

function PlaceCellScore(place field)
Input : place field ($K \times K$ 2D array)
Output : place score
$G :=$ 2D grid graph ($K \times K$)
$\mathcal{G} :=$ empty directed graph ($K \times K$)
for *edge* ($node_i \rightarrow node_j$) *in* **G do**
    if *firing rate ρ_i > firing rate ρ_j* **then**
        $\mathcal{G}$ add *edge* ($node_i \rightarrow node_j$)
    end
end
Find $node_k$ of firing rate ρ_{max}
for *$node_v$ in $\mathcal{G}$* **do**
    if *$node_v$ is not descendant of $node_k$ found with DFS(k)* **then**
        delete $node_v$ from $\mathcal{G}$
    end
end
conn. components = sum of all nodes' firing rates in $\mathcal{G}$
total components = sum of all nodes' firing rates in G
place score $= \gamma \dfrac{\text{conn. components}}{\text{total components}}$
† γ is discount factor, determined by connected component size
return place score

<center>Algorithm 1: Pseudo code for calculating place cell score metric</center>

The place field in Algorithm 1 is measured as following procedure:

1. During a random walk simulation, the activation value of a neuron at node i, where the agent is located, is measured every 65 steps. Let's say this value is a_i.

2. Every time the agent visits node i again, value a_i is added cumulatively to the recorded value; $A_i += a_i$ such that A_i is the cumulative activation value at node i. We assume the initial value of A_i is zero.

3. After the random walk is done, A_i divided by the length of the random walk trajectory is the firing rate ρ_i at node i of the neuron (place field $\in \mathbb{R}^{K \times K}$).

In our place cell evaluation experiment, the length of the random walk is 10^5 and $K = 11$; the evaluate map is one of the training maps.

A.2 Derivation of NMDAR nonlinearity from the molecular level chemical interaction

Here, we describe the NMDAR nonlinear dynamics from chemical interaction between Mg^{2+} and NMDAR following previous literature [S12–S14]. At the molecular level, one Mg^{2+} ion binds to one NMDAR receptor when opening the NMDAR channel. Thus, the chemical equation of binding reaction between Mg^{2+} ion and NMDAR receptor, R, can be described as

$$Mg^{2+} + R \rightleftharpoons Mg^{2+}R. \tag{S2}$$

From this chemical equation, the equilibrium constant K is given by

$$K = \frac{[Mg^{2+}R]}{[Mg^{2+}][R]}. \tag{S3}$$

Thus, dissociation constant K_D, which correspond to Mg^{2+} dissociation from NMDAR, follows

$$K_D = K^{-1} = \frac{[Mg^{2+}][R]}{[Mg^{2+}R]}, \tag{S4}$$

in which [R] and [Mg^{2+}R] are the free and Mg^{2+}-bound NMDARs respectively. The fraction of opened NMDAR channels (number of unbound NMDAR over a number of total NMDAR) at equilibrium follows,

$$\mathbf{p} = \frac{[\text{R}]}{[\text{R}] + [\text{Mg}^{2+}\text{R}]} \tag{S5}$$
$$= \frac{1}{1 + [\text{Mg}^{2+}]/K_D}$$

Experimentally, the voltage-dependent dynamics of K_D has been described as following equation by Ascher and Nowak [S15]

$$K_D = K_{\text{Mg}^{2+}} e^{\beta V}, \tag{S6}$$

where, V is membrane voltage, β is a temperature constant and $K_{\text{Mg}^{2+}}$ is a dissociation constant at $V = 0$. If Eq. S6 is substituted into Eq. S5, voltage-dependent open fraction of NMDAR can be expressed as follows:

$$\mathbf{p}(V) = \frac{1}{1 + \frac{[\text{Mg}^{2+}]}{K_{\text{Mg}^{2+}}} e^{-\beta V}}. \tag{S7}$$
$$= \frac{1}{1 + \alpha e^{-\beta V}}.$$

in which $\alpha = [\text{Mg}^{2+}]/K_{\text{Mg}^{2+}}$, the parameter determined by the $[\text{Mg}^{2+}]$. Given the voltage-dependent open fraction of NMDAR, $\mathbf{p}(V)$, and NMDAR's maximal conductance, g_{max}, then voltage-dependent NMDAR conductance $g(V)$ can be described as

$$g(V) = g_{\text{max}}\mathbf{p}(V) \tag{S8}$$

Given $g(V)$, and driving force, $V - V_r$, and current I, they have a relationship of $I = (V - V_r)g(V)$, in which V_r is reversal potential (the value of membrane potential above which current inverts the direction). As experimental investigations on the physiological reversal potential of NMDAR to be $V_r = 0$ [S16–S18], $I = Vg(V)$. Then, the normalized NMDAR current $I_{\text{norm}} = I/g_{\text{max}}$ follows:

$$I_{\text{norm}} = V\mathbf{p}(V) \tag{S9}$$

From Eq. S9 and previous electrophysiological experimental results [S19], we reconstruct IV curve in (Fig. 1a, right top).

A.3 NMDAR-inspired nonlinear activation function

Here, we propose an NMDAR-inspired nonlinear activation function from the nonlinear dynamics of the NMDAR-IV curve. If we consider the nonlinear IV curve of NMDAR (Eq. S9) as a nonlinear mapping function, ϕ, the membrane voltage, V, can be viewed as an input, x, and normalized NMDAR current, I_{norm}, as an output, $\phi(x)$. Therefore, we can rewrite the nonlinear mapping function ϕ as follows

$$\phi(x) = x\mathbf{p}(x). \tag{S10}$$

We define the NMDAR-inspired activation function as a nonlinear mapping function, $\text{NMDA}(x) := \phi(x)$. By substituting Eq. S7 into Eq. S10, we show the generalized expression of $\text{NMDA}(x)$ equation with α and β parameters as follows:

$$\text{NMDA}_{\alpha,\beta}(x) = x\mathbf{p}_{\alpha,\beta}(x) \tag{S11}$$
$$= \frac{x}{1 + \alpha e^{-\beta x}}.$$

Given $\alpha = 1$ and $\beta = 1$, $\mathbf{p}(x)$ is identical to the sigmoid function, $\sigma(x) = 1/(1 + e^{-x})$. This particular case of α and β leads to $x\sigma(x)$, Sigmoid Linear Unit (SiLU) activation function [S20]. In the case of $\alpha = 1$ and $\beta = 1.702$, $x\sigma(1.702x)$ corresponds to the GELU activation function, which is popular in transformer models [S21, S22, S21]. Ramachandran et al. [S23] introduced the swish activation function, $x\sigma(\beta x)$, which is a generalized form of GELU and SiLU. They demonstrated
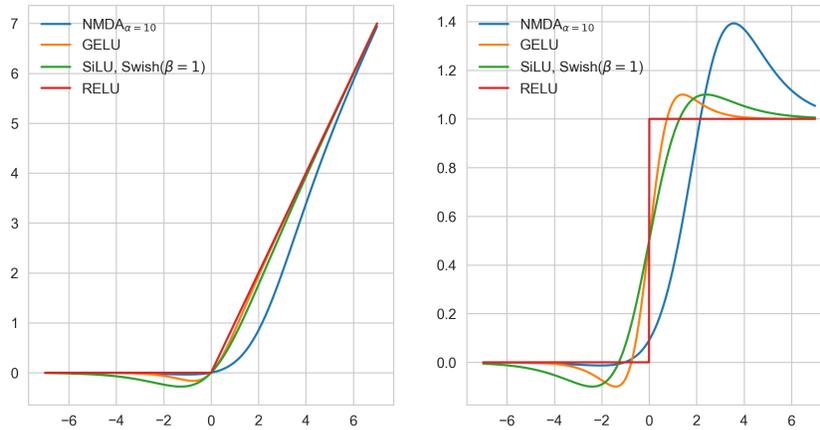
4

Figure S2: Comparison of common activation functions (left) and their derivatives (right) with $\text{NMDA}_{\alpha,\beta}$.

that when $\beta \to \infty$, the activation function resembles RELU. We summarized these four activation functions by comparing them with our $\text{NMDA}_{\alpha,\beta}(x)$ in Table 1 and Fig. S2.

In contrast to the extensive research on β in $\text{NMDA}_{\alpha,\beta}(x)$, α, the Mg^{2+}-gating component, is not explored. For this reason, we focused on the parameter α over β, and investigated $\text{NMDA}_{\alpha}(X)$. It is interesting to note that the Swish function was originally proposed as a self-gating function, inspired by the use of the sigmoid function as a gating of information flow in the long short-term memory (LSTM) network [S24]. In contrast, our activation function $\text{NMDA}(x)$ is inspired by the physical Mg^{2+}-gating mechanism that occurs at the real biological synapses. These shared mechanisms of self-gating in artificial models and biological observations raise the interesting possibility that NMDAR is a neural substrate of nonlinear activation function in the brain.

A.4 Detailed description of task design and definition of short-term working memory and long-term reference memory

Our task is based on a widely employed neuroscience experiment for two separate memory systems: working memory and reference memory [S25, S26]. Errors in working memory are measured by within-trial error, whereas errors in reference memory are measured by across-trial error. The training phase and the test phase alternate at each trial. In the test phase, the unvisited place prediction error and visited place prediction error for the familiar map and the novel map, respectively, are measured. The memory of a relatively recent experience can be defined as *short-term working memory* (STWM), and the memory of relatively old experience can be defined as *long-term reference memory* (LTRM). Within trial visited place prediction measures relatively short-term experience for our task. On the other hand, the across-trial unvisited place prediction task in the familiar map measures the relatively long-term experience. Measuring unvisited place prediction error in the novel map will establish a baseline of chance-level accuracy; above this baseline, the formation of long-term memory can be observed (Fig. S3).
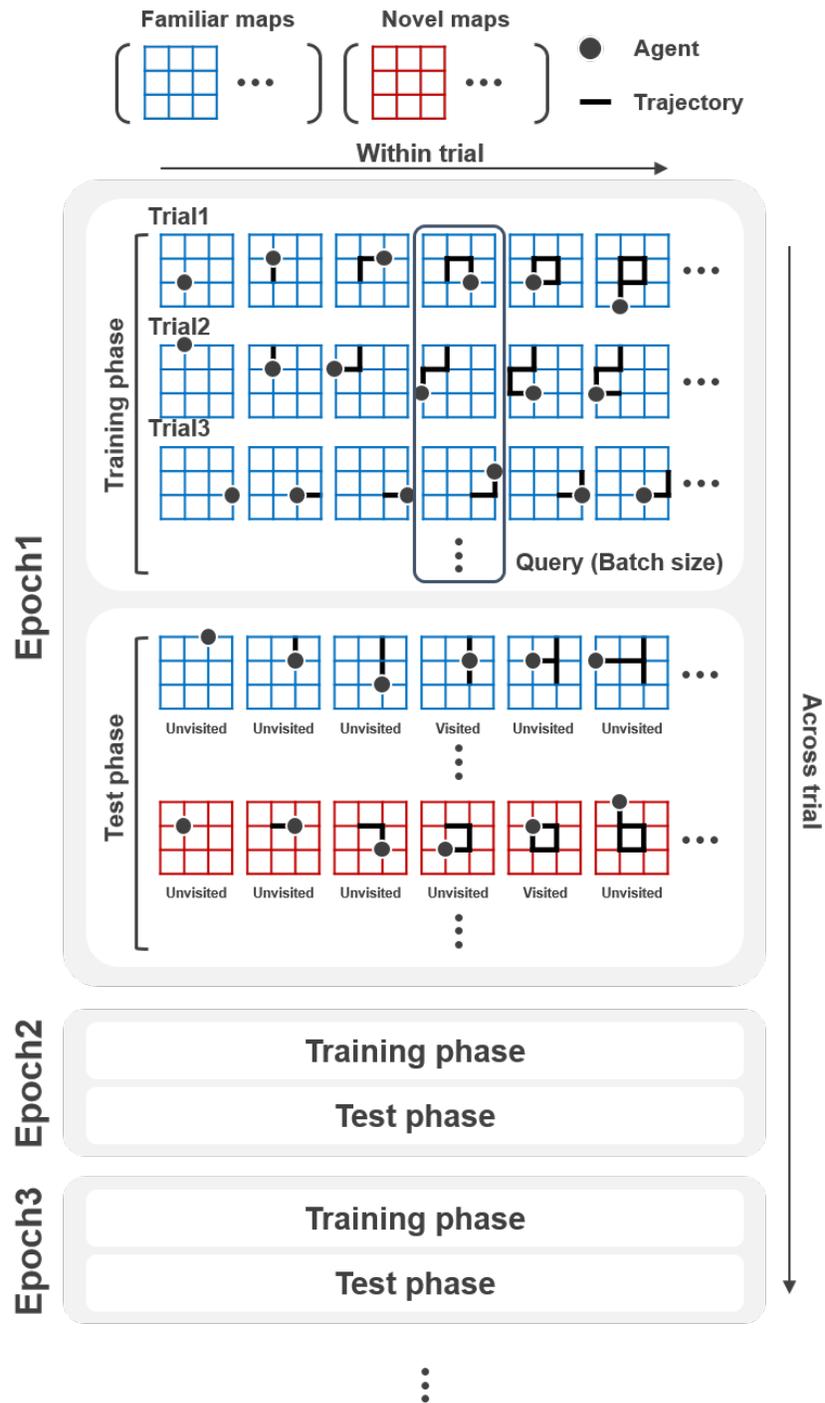
Figure S3: Detailed task design of working and reference memory evaluation. At each random walk step, a batch is created (which is then used in the backpropagation step). The batch size is 512 since there are 512 parallel random walkers in use. Note that at each trial the agent randomly selects a map from training maps (familiar maps), the initial position of the agent is random, and the agent does a random walk.

A.5 Non-recurrent positional embeddings and prediction errors on the node visited for the first time
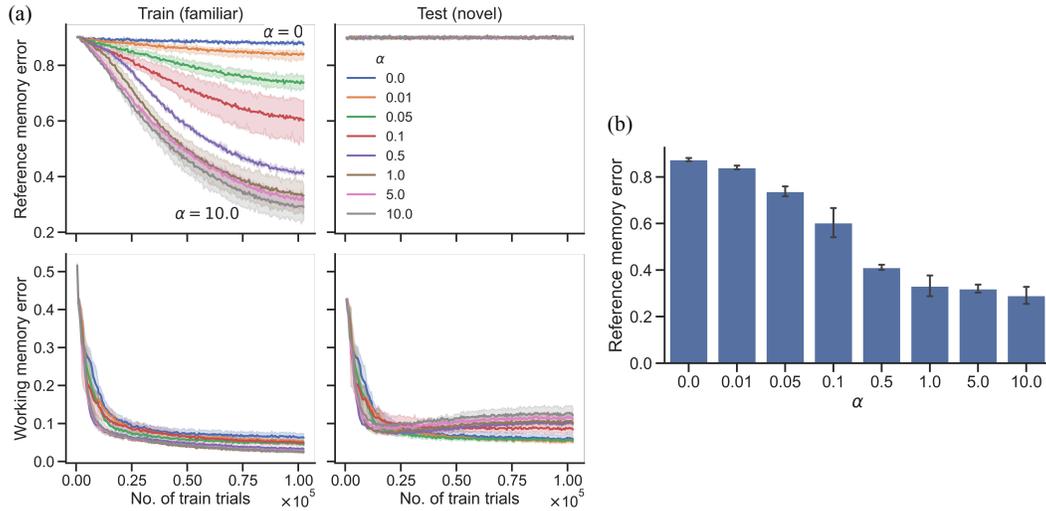
Figure S4: Experiment with non-recurrent positional embeddings. (a) Reference and working memory errors over training trials for training (familiar) maps and testing (novel) maps for $N = 32$ where N is the number of training maps. (b) Reference memory errors evaluated on training maps over different values of α in NMDA$_\alpha$ for $N = 32$. Error bars and shaded areas represent the standard deviation of errors from three independently trained models.

Non-recurrent positional embeddings We tested the non-recurrent positional embedding by substituting the recurrent positional embedding $\mathbf{e}_t$ with the action embedding $A(a_t)$, where A is the embedding layer and a_t is the action at step t. Compared to Fig. S4a, the result demonstrates a significant increase in working memory error and reference memory error (Fig. 3 vs. Fig. S4). Nonetheless, the model's behavior is comparable to the trend of decreasing reference memory error with increasing α of NMDA$_\alpha$ (see Fig. S4b).

Moreover, we found that place cells do emerge in the feedforward layers for this standard positional embedding method (see Fig. S5). It is worth noting that these place cell scores are noticeably lower compared to those obtained through the recurrent positional embedding. Despite the hindered performance in working memory and reference memory, a strong correlation persists between the place cell scores in FFNs and the reference memory errors (see Fig. S5c and S5d).

Prediction errors on the node visited for the first time We compared unvisited node prediction error (unvisited within context window, in this case, 64 steps) versus first visited node prediction error (unvisited for within a trial). As shown in Fig. S6, the prediction error results for the first visited node do not differ from the reference memory error results.

These results strongly support the conclusions that (1) while the path-integrated information from recurrent positional embedding is important for learning the spatial structure of the map, this information is not used in predicting the unvisited node, and (2) the reference memory is used for predicting an unvisited node on a familiar map.
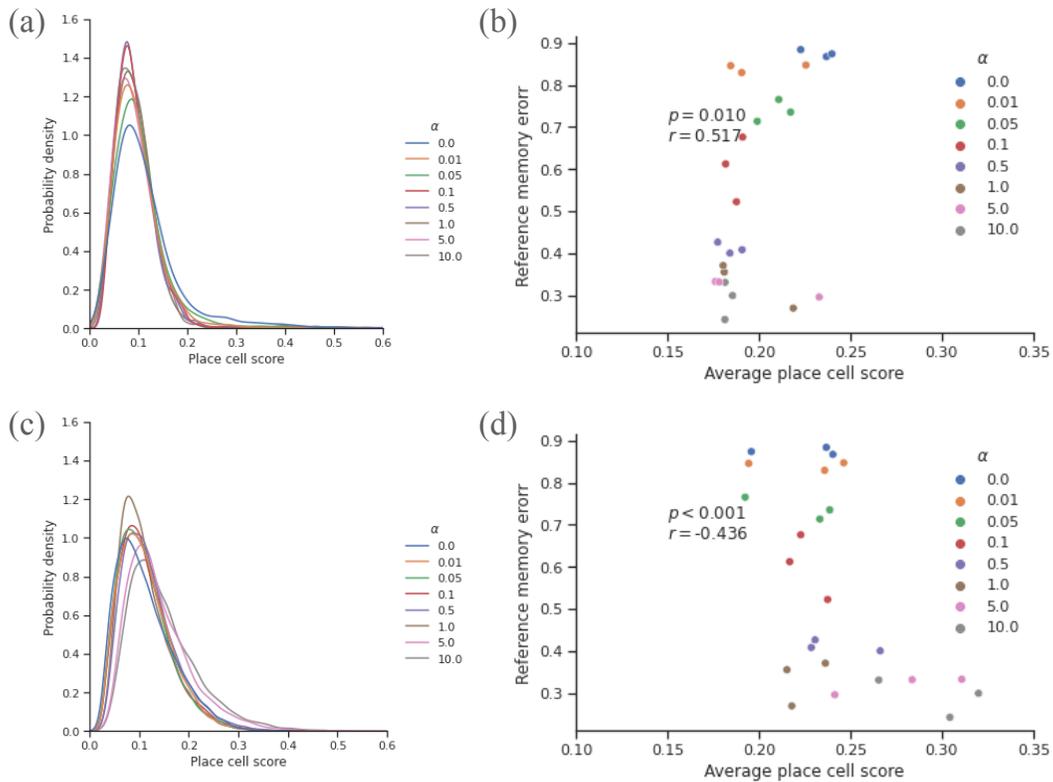
Figure S5: Place cell score analysis for non-recurrent positional embeddings for $N = 32$. Place cell score distributions with varying α in self-attention layers (a) and feed-forward layers (c). Scatter plot of average place cell scores and reference memory errors for self-attention layers (b) and feed-forward layers (d). r and p denote Spearman's rank correlation coefficient and significance score, respectively.
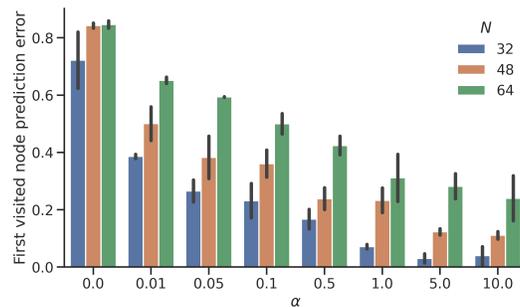

Figure S6: First visited node prediction error evaluated on training maps over different values of α in NMDA$_\alpha$ for $N = 32, 48$, and 64. Error bars and shaded areas represent the standard deviation of errors from three independently trained models.

A.6 Statistical significance between the NMDA and GELU activation functions & learning rate effect: is it an under-trained effect?

In this section, we address concerns about the NMDA style activation with large α being a trivial effect for the emergence of place cells and the improvement of learning speed. We provide additional evidence to support our claims and demonstrate that the observed effects are not trivial or due to under-training.

Statistical test across activation functions   Our primary claim regarding the effectiveness of NMDA-style activation is based on the results shown in Fig. 3. To support our claim, we conducted a

statistical across various activation functions. Our findings revealed a significant difference between $NMDA_\alpha$ and other activation functions, indicating that the NMDA style activation is beneficial for the emergence of place cells and the improvement of learning speed (see Table S7).
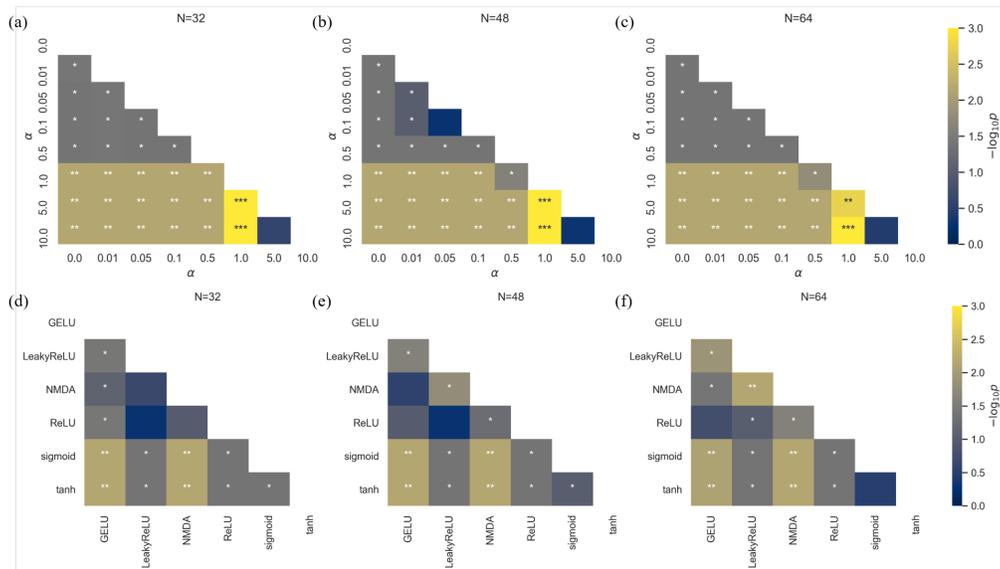

Figure S7: Statistical test result across various activation functions with different numbers of maps (N) related to Figure 3. (a-c) Mann-Whitney test across various α of $NMDA_\alpha$. (d-f) Mann-Whitney test across various activation functions. The sample size (or trials) used for tests ranges from 3 to 11. Colors represent p-values and asterisks indicates statistical significances: * $0.01<p<0.1$; ** $0.001<p<0.01$; *** $p<0.001$. $\alpha = 10$ for NMDA and $\alpha = 0.01$ for LeakyReLU.

Gradient around $x = 0$ and learning speed The performance gain at large alpha values might be due to an increased gradient around $x = 0$. However, the gradient around $x = 0$ at $\alpha = 1$ (green) is actually larger than $\alpha = 10$ (blue) in Fig. S2. Therefore, the performance gain at large alpha values is not due to an increased effective learning speed around $x = 0$. Instead, it may be attributed to the activation function's ability to better capture nonlinear relationships.

Learning rate effect on reference memory error To clarify that our result is not simply due to the under-trained model, we performed experiments with varying learning rates: 0.001, 0.0005, 0.0003, 0.0002, and 0.0001. As can be seen in Fig. S8, a larger learning rate at low alpha values is not helpful for reducing reference memory error. It is worth noting that all models' working memory error is almost zero. Larger learning rates lead to faster decrease in working memory error, but at high learning rates, reference memory error stays around the chance level of 0.9, i.e., no reference memory emerges. Additionally, Fig. S9 shows that higher place cell scores correspond to lower reference memory errors. Note that we employed a linear decay learning rate schedule (from the starting learning rate to 0) and the starting learning rate used in our manuscript is 0.0001.

In conclusion, our additional experiments and analyses demonstrate that the NMDA style activation with large α plays a critical role in the emergence of place cells and the improvement of reference memory, and it is not a trivial effect or an under-trained effect.
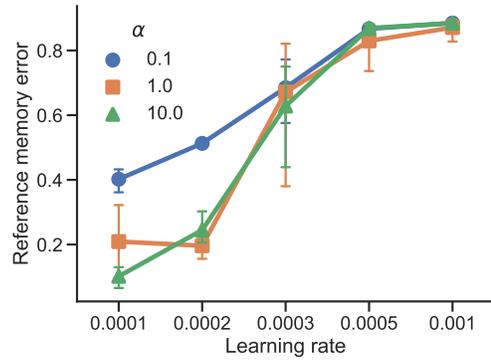
Figure S8: Familiar map reference memory error results with five different learning rates, where $\alpha \in \{0.1, 1.0, 10.0\}$, for a total of 64 training maps ($N = 64$). The error bars represent the standard deviation of errors from three independently trained models.
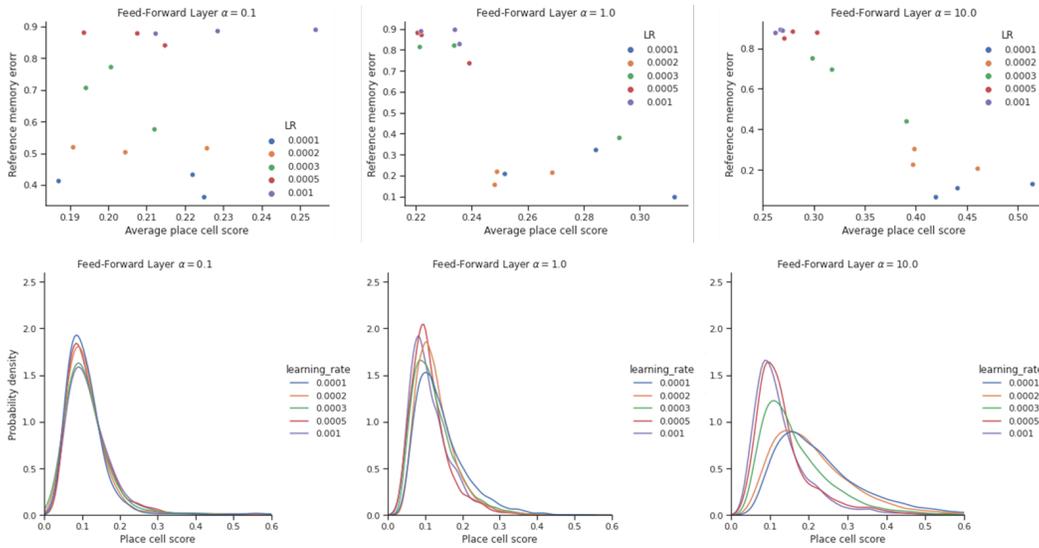

Figure S9: (top) Scatter plot between reference memory error and average place score; (bottom) Distribution of place cell scores with five different learning rates, where $\alpha \in \{0.1, 1.0, 10.0\}$ (shown at the top of each figure), for a total of 64 training maps ($N = 64$).

A.7   Consequent of changing nonlinear dynamics in Leaky ReLU activation function

Here, we investigated the consequence of changing nonlinearity with an activation function other than NMDA$_\alpha$. We chose LeakyReLU ($\max(0, x) + \alpha \min(0, x)$) activation function to compare with NMDA$_\alpha$. Regarding LeakyReLU, $\alpha = 1$ of LeakyReLU also leads to a fully linear activation function similar to $\alpha = 0$ of NMDA$_\alpha$. Compared to NMDA$_{\alpha=10}$, LeakyReLU exhibits a lower average place score in the allowed range of α, indicating that NMDA$_\alpha$ performs better in terms of robust place cell emergence (see Fig. S10)
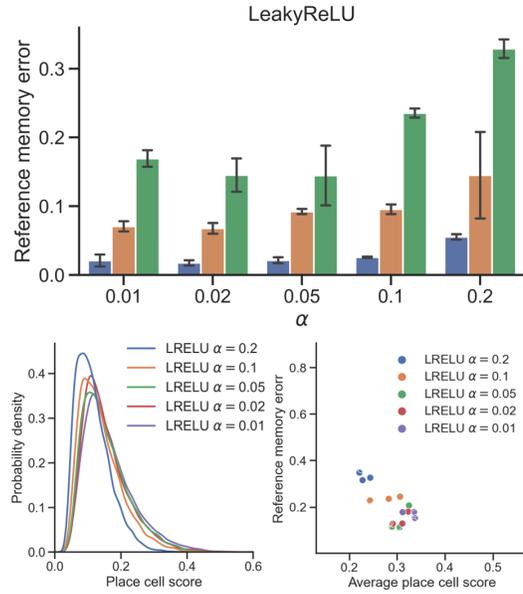
Figure S10: Evaluation of reference memory error in LeakyReLU (LRELU) while modulating α(top) and relationship of average place cell score and reference memory error (bottom).

A.8 Transformer as a memory consolidation model and its biological plausibility

Next, we examined the biologically inspired NMDA$_\alpha$ activation function in the feed-forward layer of the transformer and its role in memory formation and place cell representation. We show that modulating α corresponds to a change in extracellular [Mg^{2+}], by deriving the nonlinear activation function from the real NMDAR nonlinear IV curve. The reconstructed real nonlinear IV curve is shown in Fig. 1a (top right panel).

The modulation of α selectively affects the formation of long-term reference memory (i.e., prediction of unvisited places across trials) while leaving the formation of short-term working memory (i.e., prediction of unvisited places within trials) intact. This result suggests that short-term working memory and long-term reference memory are physically stored in separate structures: the self-attention layer and the feed-forward layer. In psychology, a multi-store model for the animal brain that comprises short-term memory and long-term memory was historically suggested in [S27]. In this biological model, sensory inputs are stored in short-term memory systems via attention, and some are transferred to a long-term memory system while others quickly disappear.

In neuroscience, the transfer of short-term memory into a long-term system is called *memory consolidation* [S28]. Animal studies have demonstrated that the CA1 region of the hippocampus is essential for memory consolidation [S29, S30]. In hippocampal CA1, the postsynaptic NMDA receptor mediates synaptic plasticity, and the selective perturbation of these receptors leads to impairment in long-term memory formation [S31, S30]. Later research revealed that Mg^{2+}-gating of NMDA receptors modulates the formation of long-term memory [S32, S33]. These observations imply that the nonlinear dynamics of NMDA receptors in CA1 are critical for consolidating short-term memory into long-term memory S11.

On the basis of a previous link between the hippocampus and the transformer model, we hypothesize that the latter is a model for memory consolidation. Given the resemblance of the GELU nonlinear activation function and CA1 NMDAR nonlinear IV curve, we assumed that the GELU activation function serves as a key component that links short-term working memory and long-term reference memory. Our experimental results indicate that that the formation of long-term reference memory is impaired when the activation function is completely linear (corresponding to no Mg^{2+}). In contrast, increasing α (which corresponds to an increase in Mg^{2+} level) revealed that for long-term reference memory, our model outperforms other activation functions (e.g., RELU, GELU, LRELU, Sigmoid, Tanh). Based on these similarities between hippocampal memory consolidation and our results, we propose that transformer is an effective memory consolidation model.
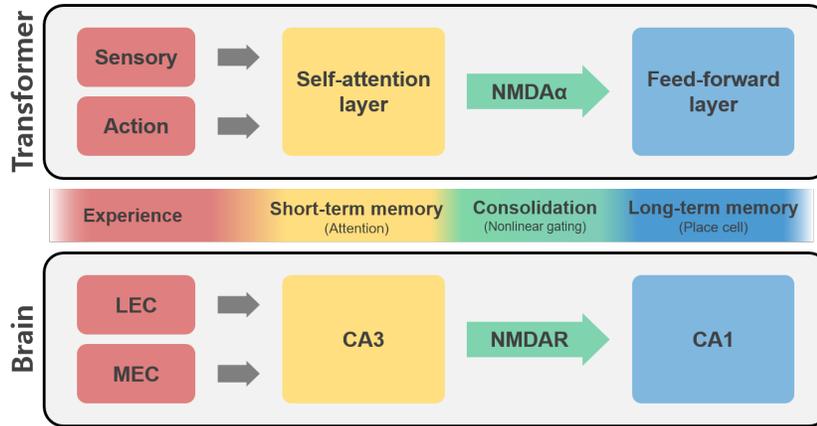
Figure S11: Schematic illustration of the transformer as a memory consolidation model of hippocampus via NMDAR nonlinearity. Top, transformer with $NMDA_\alpha$ nonlinear activation function.Bottom, summarized brain model with NMDA receptor mediated long-term memory formation. Colors represent the corresponding stage of the memory consolidation process.

In addition to the performance gain in long-term memory formation with $NMDA_\alpha$, we found that modulating the α affects the emergences of place cells in the feed-forward layer and conclude that there is a significant correlation between place cell score and long-term reference memory formation. Our results align with previous biological findings that perturbation of CA1 NMDARs leads to impairment in both place cell representation and long-term memory formation [S31, S34, S35, S29]. These similarities support the idea that place cells are the neural correlates of long-term spatial memories. Altogether, our findings suggest the exciting possibility that the nonlinear IV curve of NMDAR in the hippocampal CA1 is a neural substrate of nonlinear activation function in the brain.

A.9 GPT2 and Vision Transformer with NMDA activation functions

In this section, we present the detailed experimental setup and results for applying the proposed $NMDA_\alpha$ activation function to the GPT2 [S36] and Vision Transformer (ViT) models.

A.9.1 Language Modeling with GPT2

We replaced the GELU activation function in the feed-forward networks (FFNs) of the GPT2 model (model size: 124M) with the $NMDA_{\alpha=10}$ activation function and conducted experiments on the Open-WebText dataset [S37]. We employed the GPT2 training PyTorch code from `karpathy/nanoGPT` GitHub repository [S38]. The application of $NMDA_{\alpha=10}$ showed a slight increase in performance, as illustrated in Fig. S12.

A.9.2 Image Classification Tasks with ViT

We conducted experiments on image classification tasks using the ViT model, and the results were examined on the CIFAR-100 and TinyImageNet datasets. Three trials were conducted for each condition, and an increasing, though statistically insignificant, tendency in performance was observed. The top-1 test accuracies for these datasets are provided in Table S1.

Table S1: Top-1 test accuracies with ViT model on CIFAR-100 and TinyImageNet datasets.

Dataset	GELU	$NMDA_{\alpha=10}$	$NMDA_{\alpha=0}$
CIFAR100	69.92 ± 0.34	70.27 ± 0.47	49.91 ± 0.58
TinyImNet	54.90 ± 0.45	55.72 ± 0.03	40.36 ± 0.52

Hyperparameter Consideration It is important to note that we did not scan a range of hyperparameters for either NMDA or the GELU-based model. The optimal hyperparameters for NMDA might differ from those for the GELU-based model.
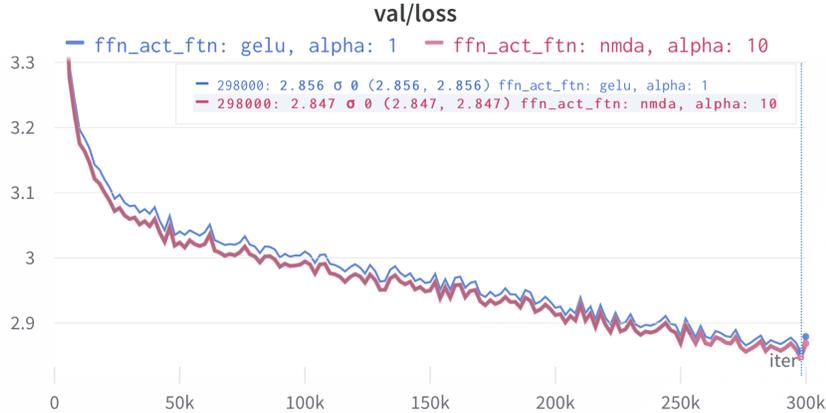
Figure S12: Validation loss curve over training iteration for GPT2 model on the OpenWebText dataset. The blue (red) line represents the GPT2 with GELU (NMDA$_{\alpha=10}$) function.

Future Works Although these results are promising, they are not conclusive in stating that NMDA generally outperforms GELU for various real-world tasks. Future works will include evaluating our NMDA function on different model sizes of GPT or other language models, as well as evaluating the pre-trained models on downstream NLP tasks such as reading comprehension, question answering, common sense, MMLU, and BIG-bench.

A.10   Sparsity of Activation

In this section, we present more intuitive evidence to support the effectiveness of our proposed nonlinearity, based on the sparsity of activation in FFNs.

Mathematical Intuition Given the NMDA$_\alpha$ activation function, NMDA$_\alpha(x) = x/(1 + \alpha e^{-x})$, we can rewrite this function as follows:

$$\text{NMDA}_{\alpha(x)} = x \cdot \frac{1}{1 + e^{-(x-c)}}, \quad \text{where } \alpha = e^c. \tag{S12}$$

In the given expression, an increase in α results in shifting the sigmoid function towards the positive direction by increasing c. This shift can be understood as an increase in the threshold of information gating, which is mediated by the sigmoid function. Consequently, an increase in α might enhance the sparsity of the NMDA activation.

Empirical Results To validate the aforementioned mathematical intuition, we conducted measurements on the sparsity of the activities in the feed-forward layer by calculating the Gini index [S39] for each input sequence. The Gini index is defined as follows:

$$G = \frac{\sum_{i=1}^{K} \sum_{j=1}^{K} |x_i - x_j|}{2K^2 \bar{x}}, \tag{S13}$$

where x_i is the i-th neuron's activation value and $\bar{x} = \sum_{i=0}^{K} x_i / K$ is the mean of the activation values in the feed-forward layer ($K = 2048$ is the total number of neurons in a feed-forward layer).

The Gini index ranges from 0 to 1. When only a few neurons have very high activation values and others are small, the Gini index is close to 1; on the other hand, when most neurons have homogeneous activation values, this value is close to 0 (1 = absolute sparsity, 0 = all activations equal).

According to Fig. S13a, when the value of α in NMDA$_\alpha(x)$ is increased, the Gini index also increases. This increase signifies that the population activities become more heterogeneous, leading to a heavy-tailed distribution. Furthermore, the Gini index of NMDA with an α value of 10 is greater than that of the GELU activation function (indicated by the dashed line). These findings imply that NMDA$_\alpha(x)$ could potentially improve long-term memory formation by promoting sparsity in the activations of the feed-forward layer.
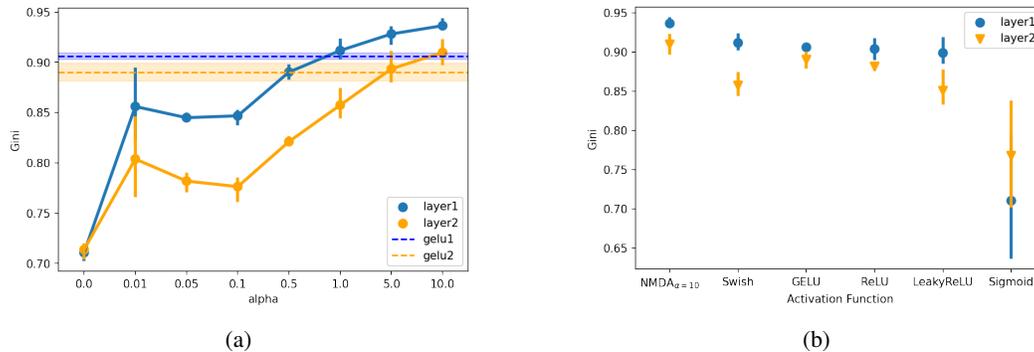
Figure S13: (a) Comparison of Gini index for different values of α in NMDA_α. (b) Comparison of Gini index between different nonlinear activation functions.

Previous studies have explored the generalization performance and sparsity in overparameterized models [S40]. Although the mechanism behind the emergence of sparse representation in large models is not fully understood, it is worth noting that overparameterized models converge to simpler models than those with dense representations.

The high sparsity properties of the NMDA function may contribute to the increased score of place cells. In the case of Transformers, enforcing neuron activation sparsity in MLPs has been found to improve the interpretability or selectivity of a higher percentage of neurons [S41]. This evidence could potentially explain the higher place cell score observed when the alpha value is increased. We observed that NMDA with $\alpha = 10$ exhibits the highest Gini index among different activation functions. See Fig. S13a for a comparison of Gini index among activation functions.

References

[S1] John O'Keefe. Place units in the hippocampus of the freely moving rat. *Experimental neurology*, 51(1):78–109, 1976.

[S2] Matthew A Wilson and Bruce L McNaughton. Dynamics of the hippocampal ensemble code for space. *Science*, 261(5124):1055–1058, 1993.

[S3] John O'keefe and Lynn Nadel. Précis of o'keefe & nadel's the hippocampus as a cognitive map. *Behavioral and Brain Sciences*, 2(4):487–494, 1979.

[S4] Howard Eichenbaum, Paul Dudchenko, Emma Wood, Matthew Shapiro, and Heikki Tanila. The hippocampus, memory, and place cells: is it spatial memory or a memory space? *Neuron*, 23(2):209–226, 1999.

[S5] Christopher J. Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B17JTOe0-.

[S6] Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705): 429–433, 2018.

[S7] James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

[S8] James C. R. Whittington, Joseph Warren, and Tim E.J. Behrens. Relating transformers to models and neural representations of the hippocampal formation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=B8DVo9B1YE0.

[S9] César Rennó-Costa and Adriano BL Tort. Place and grid cells in a loop: implications for memory function and spatial coding. *Journal of Neuroscience*, 37(34):8062–8076, 2017.

[S10] NM Van Strien, NLM Cappaert, and MP Witter. The anatomy of memory: an interactive overview of the parahippocampal–hippocampal network. *Nature reviews neuroscience*, 10(4): 272–282, 2009.

[S11] Dori M Grijseels, Kira Shaw, Caswell Barry, and Catherine N Hall. Choice of method of place cell classification determines the population of cells identified. *PLoS Computational Biology*, 17(7):e1008835, 2021.

[S12] Ann M Woodhull. Ionic blockage of sodium channels in nerve. *The Journal of general physiology*, 61(6):687–708, 1973.

[S13] Craig E Jahr and Charles F Stevens. Voltage dependence of nmda-activated macroscopic conductances predicted by single-channel kinetics. *Journal of Neuroscience*, 10(9):3178–3182, 1990.

[S14] MICHAEL Perouansky and YOEL Yaari. Kinetic properties of NMDA receptor-mediated synaptic currents in rat hippocampal pyramidal cells versus interneurones. *The Journal of Physiology*, 465(1):223–244, 1993.

[S15] Philippe Ascher and Linda Nowak. The role of divalent cations in the N-methyl-D-aspartate responses of mouse central neurones in culture. *The Journal of physiology*, 399(1):247–266, 1988.

[S16] ML Mayer and GL Westbrook. Permeation and block of N-methyl-D-aspartic acid receptor channels by divalent cations in mouse cultured central neurones. *The Journal of physiology*, 394(1):501–527, 1987.

[S17] Tomomi Ichinose, Shun Yu, Xue Qing Wang, and Shan Ping Yu. Ca2+-independent, but voltage-and activity-dependent regulation of the NMDA receptor outward k+ current in mouse cortical neurons. *The Journal of physiology*, 551(2):403–417, 2003.

[S18] G Kohr, Yves De Koninck, and Istvan Mody. Properties of nmda receptor channels in neurons acutely isolated from epileptic (kindled) rats. *Journal of Neuroscience*, 13(8):3612–3627, 1993.

[S19] Eilon D Kirson, Claudia Schirra, Arthur Konnerth, and Yoel Yaari. Early postnatal switch in magnesium sensitivity of nmda receptors in rat ca1 pyramidal cells. *The Journal of Physiology*, 521(Pt 1):99, 1999.

[S20] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

[S21] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[S22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[S23] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[S24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997.

[S25] David S Olton, Christine Collison, and Mary Ann Werz. Spatial memory and radial arm maze performance of rats. *Learning and motivation*, 8(3):289–314, 1977.

[S26] David S Olton, James T Becker, and Gail E Handelmann. Hippocampus, space, and memory. *Behavioral and Brain sciences*, 2(3):313–322, 1979.

[S27] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier, 1968.

[S28] James L McGaugh. Memory–a century of consolidation. *Science*, 287(5451):248–251, 2000.

[S29] Eiji Shimizu, Ya-Ping Tang, Claire Rampon, and Joe Z Tsien. Nmda receptor-dependent synaptic reinforcement as a crucial process for memory consolidation. *Science*, 290(5494): 1170–1174, 2000.

[S30] Miguel Remondes and Erin M Schuman. Role for a cortical input to hippocampal area ca1 in the consolidation of a long-term memory. *Nature*, 431(7009):699–703, 2004.

[S31] Joe Z Tsien, Patricio T Huerta, and Susumu Tonegawa. The essential role of hippocampal ca1 nmda receptor–dependent synaptic plasticity in spatial memory. *Cell*, 87(7):1327–1338, 1996.

[S32] Inna Slutsky, Nashat Abumaria, Long-Jun Wu, Chao Huang, Ling Zhang, Bo Li, Xiang Zhao, Arvind Govindarajan, Ming-Gao Zhao, Min Zhuo, et al. Enhancement of learning and memory by elevating brain magnesium. *Neuron*, 65(2):165–177, 2010.

[S33] Tomoyuki Miyashita, Yoshiaki Oda, Junjiro Horiuchi, Jerry CP Yin, Takako Morimoto, and Minoru Saitoe. Mg2+ block of drosophila NMDA receptors is required for long-term memory formation and CREB-dependent gene expression. *Neuron*, 74(5):887–898, 2012.

[S34] Thomas J McHugh, Kenneth I Blum, Joe Z Tsien, Susumu Tonegawa, and Matthew A Wilson. Impaired hippocampal representation of space in ca1-specific nmdar1 knockout mice. *Cell*, 87 (7):1339–1349, 1996.

[S35] Clifford Kentros, Eric Hargreaves, Robert D Hawkins, Eric R Kandel, Matthew Shapiro, and Robert V Muller. Abolition of long-term stability of new hippocampal place cell maps by NMDA receptor blockade. *Science*, 280(5372):2121–2126, 1998.

[S36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[S37] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[S38] Andrej Karpathy. The simplest, fastest repository for training/finetuning medium-sized gpts. hhttps://github.com/karpathy/nanoGPT, 2023.

[S39] Michelle Miller, SueYeon Chung, and Kenneth D. Miller. Divisive feature normalization improves image recognition performance in alexnet. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aOX3a9q3RVV.

[S40] Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J Reddi, Ke Ye, Felix Chern, Felix Yu, Ruiqi Guo, et al. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. In *The Eleventh International Conference on Learning Representations*.

[S41] Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer ElShowk, Nicholas Joseph, Nova DasSarma, Ben Mann, Danny Hernandez, Amanda Askell, Kamal Ndousse, Andy Jones, Dawn Drain, Anna Chen, Yuntao Bai, Deep Ganguli, Liane Lovitt, Zac Hatfield-Dodds, Jackson Kernion, Tom Conerly, Shauna Kravec, Stanislav Fort, Saurav Kadavath, Josh Jacobson, Eli Tran-Johnson, Jared Kaplan, Jack Clark, Tom Brown, Sam McCandlish, Dario Amodei, and Christopher Olah. Softmax linear units. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/solu/index.html.