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†]

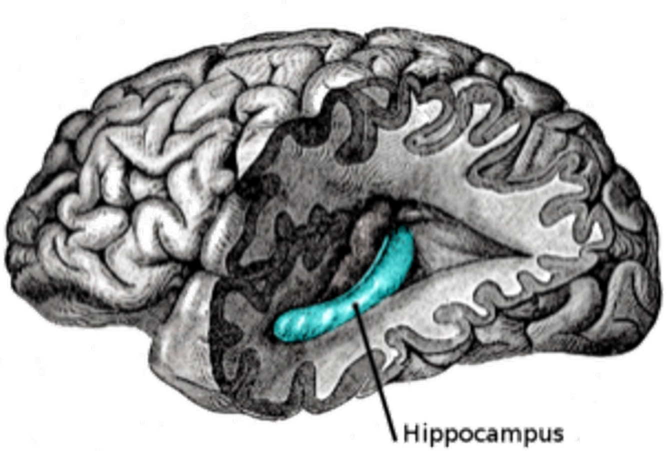


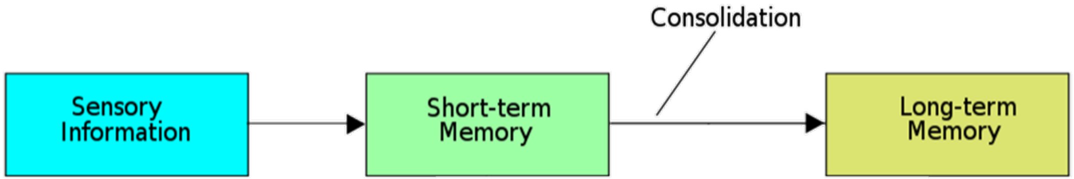

[1]Data Science Group, IBS [2]Center for Cognition and Sociality, IBS [3]School of Computing, KAIST

[*]Equal contributions; [†]Corresponding authors.

37th Conference on Neural Information Processing Systems (NeurIPS 2023)

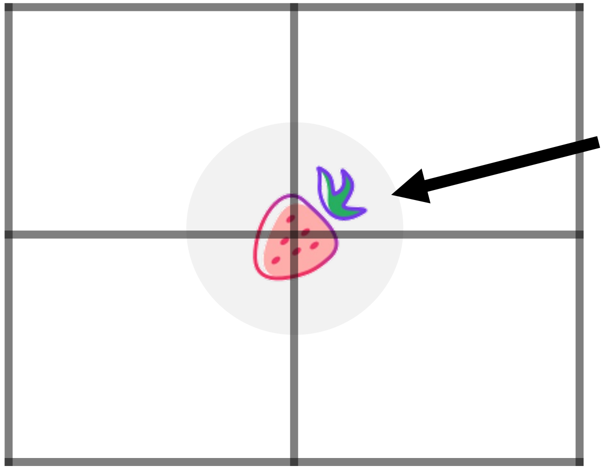
Memory formation is key in intelligent systems





“Memory Consolidation," Wikipedia, last modified April 10, 2023, https://en.wikipedia.org/wiki/Memory_consolidation

Short-term working memory use transient information

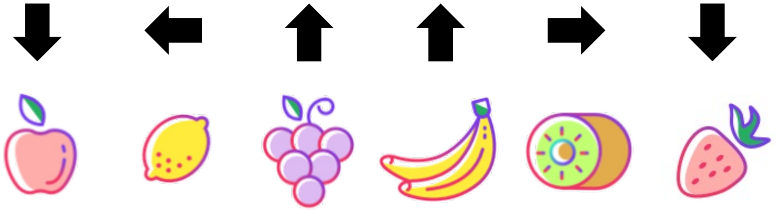


Only visible inside circle

Trial 1

Action

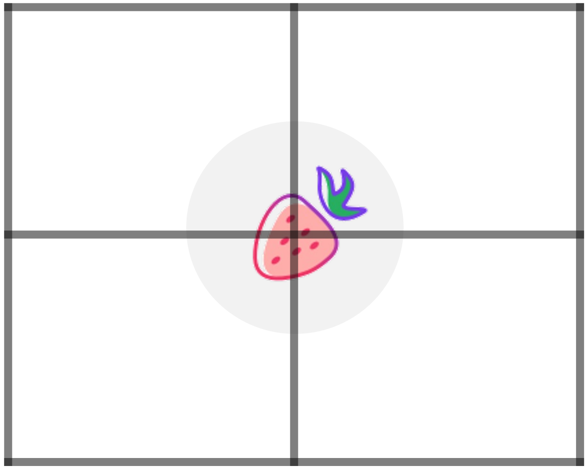
Observation



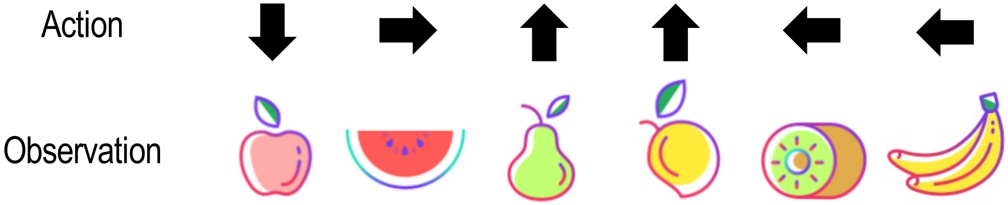
Q. What comes next?

→ Inferred by within-trial observations

Long-term reference memory retrieve previous information



Trial 2



Q. What comes next?

→ Inferred by across-trial observations

Memory consolidation is a process that transforms STWM into LTRM

short-term
working memory

use transient information
(within-trial observation)

$\xrightarrow{\text{Consolidation}}$

long-term
reference memory

retrieve previous information
(across-trial observations)

NMDA receptor in brain makes this happen

short-term
working memory

use transient information
(within-trial observation)

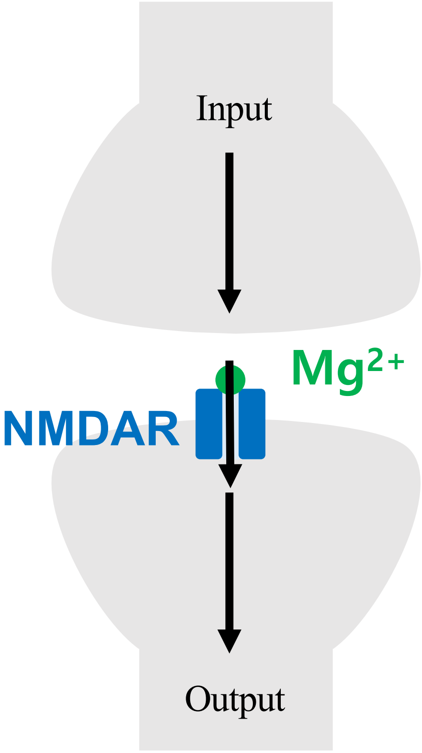
Consolidation →

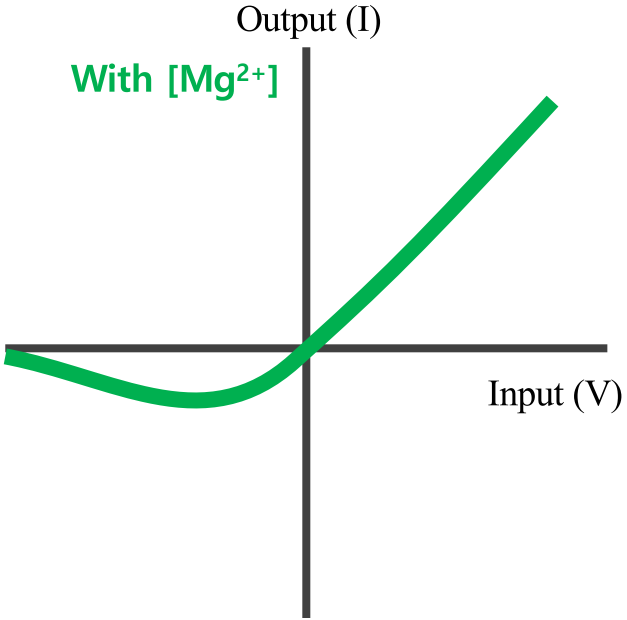
long-term
reference memory

retrieve previous information
(across-trial observations)

Shimizu et al. (2000) ***Science***
McGaugh (2000) ***Science***

NMDA receptor has nonlinearity



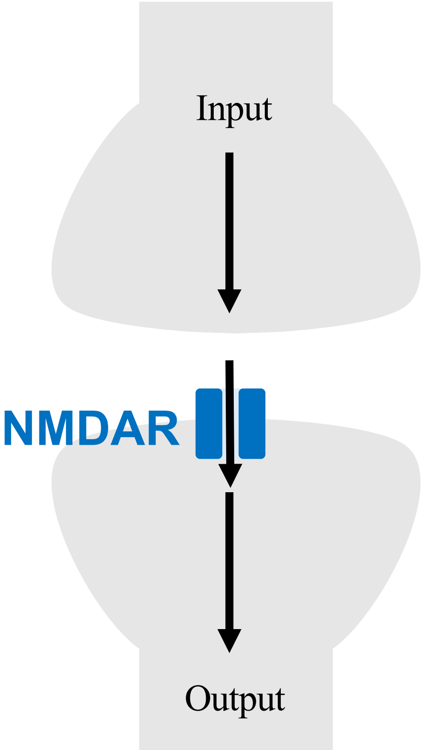


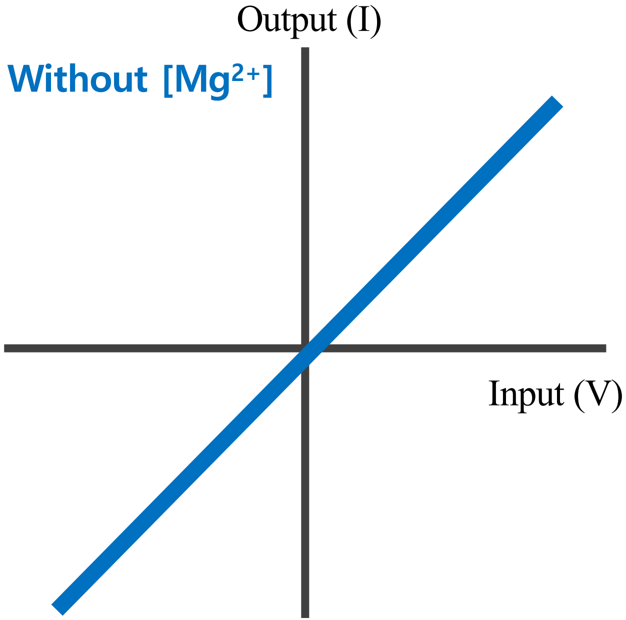
Nonlinear Dynamics
Synaptic plasticity
Memory formation

$$V \cdot \frac{1}{1 + \alpha \cdot \exp(-\beta V))}$$

$$\boxed{\frac{[\mathbf{Mg}^{2+}]}{K_{\mathbf{Mg}^{2+}}}}$$

NMDA receptor nonlinearity disappears without Magnesium ion (Mg2+)





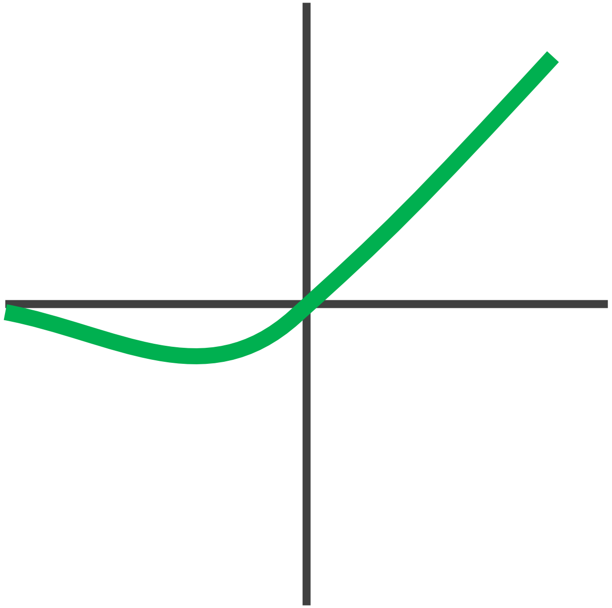
Linear Dynamics

~~Synaptic plasticity~~

~~Memory formation~~

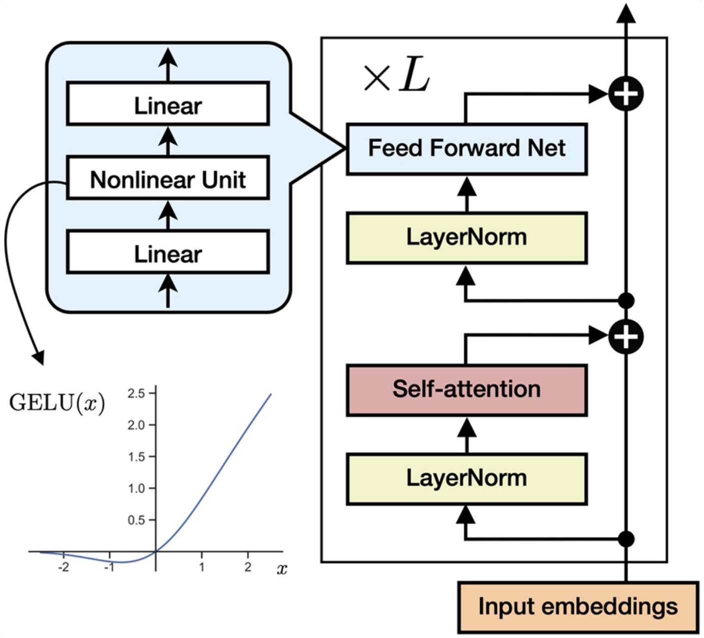
$$V \cdot \frac{1}{1 + \underbrace{0}_{\boxed{\frac{[\mathbf{Mg}^{2+}] = 0}{K_{\mathbf{Mg}^{2+}}}}} \cdot \exp(-\beta V))}$$

We observe NMDAR nonlinearity in brain resembles the nonlinear activation function in Transformer



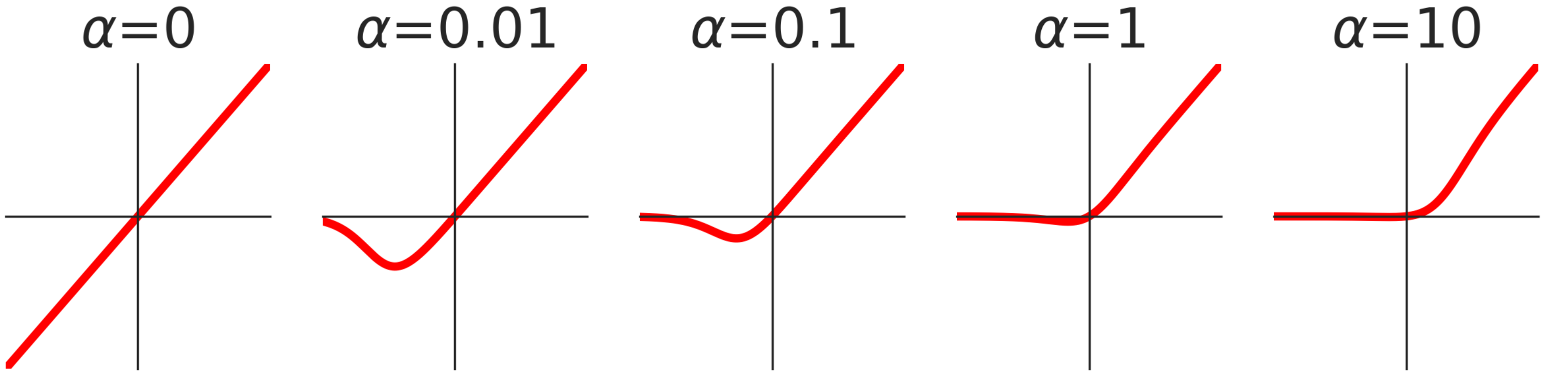
NMDAR Nonlinearity

Transformer



We propose a new activation function for transformers that mimics brain's nonlinearity (Mg^{2+} of NMDAR)

$$\mathrm{NMDA}_{\alpha}(x) = \frac{x}{1 + \alpha e^{-x}}$$



RQ 1: How would NMDAR-like nonlinearity perform in Transformer?

Memory Consolidation
STWM → LTRM

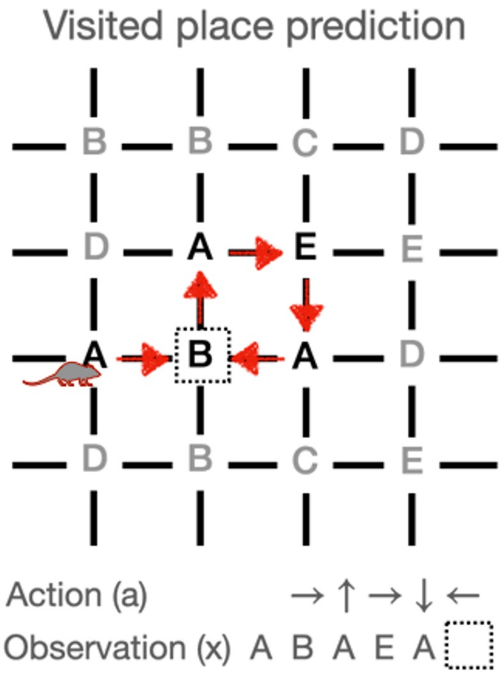
+

NMDAR
Nonlinearity

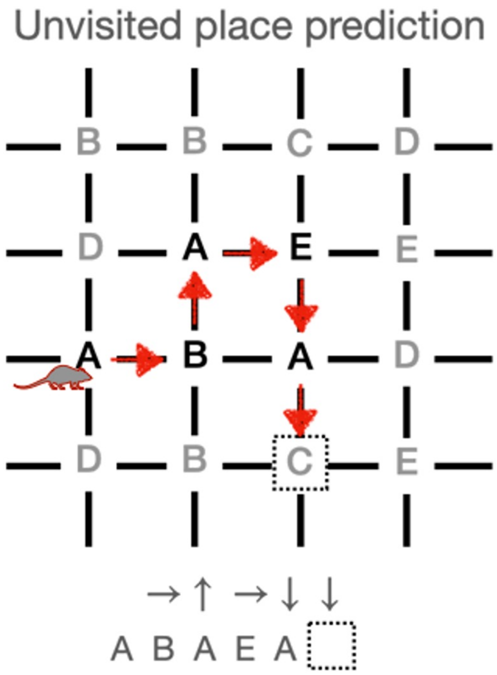
Transformer

We will test two memory functions:

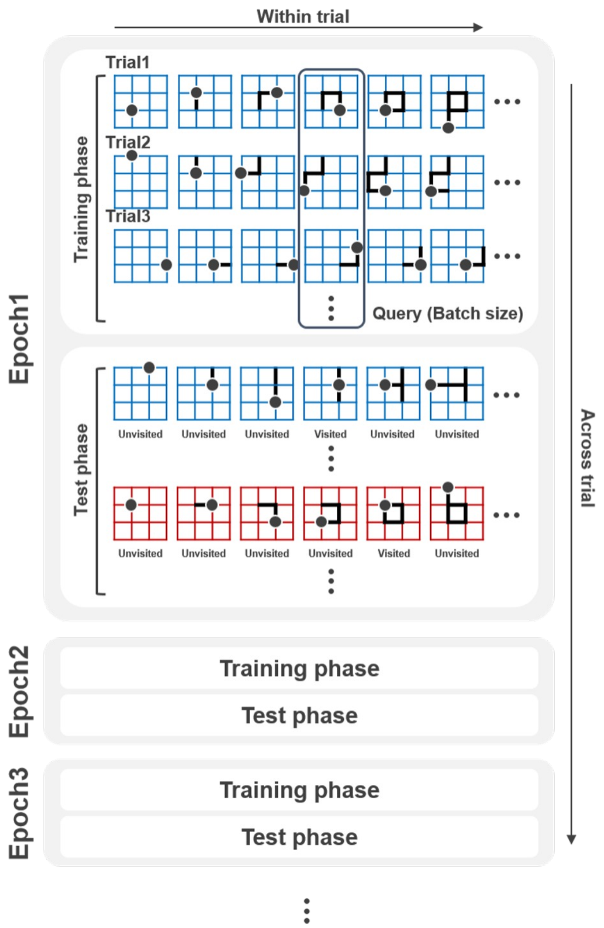
Short-term working memory

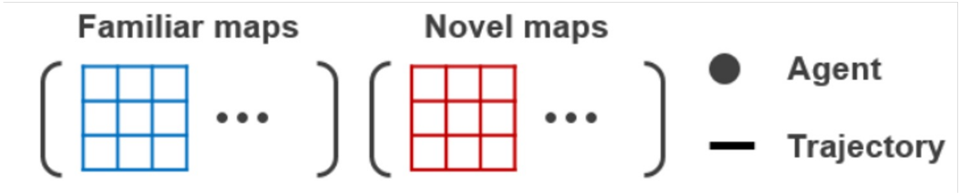


Long-term reference memory



by designing a 2D navigation task

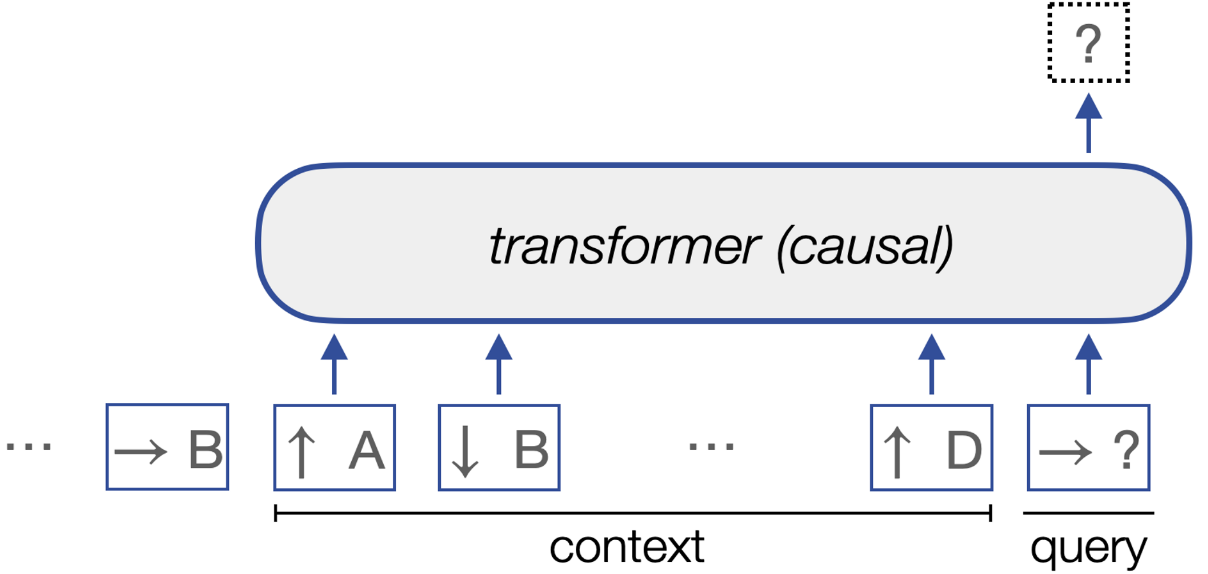




Experimental details

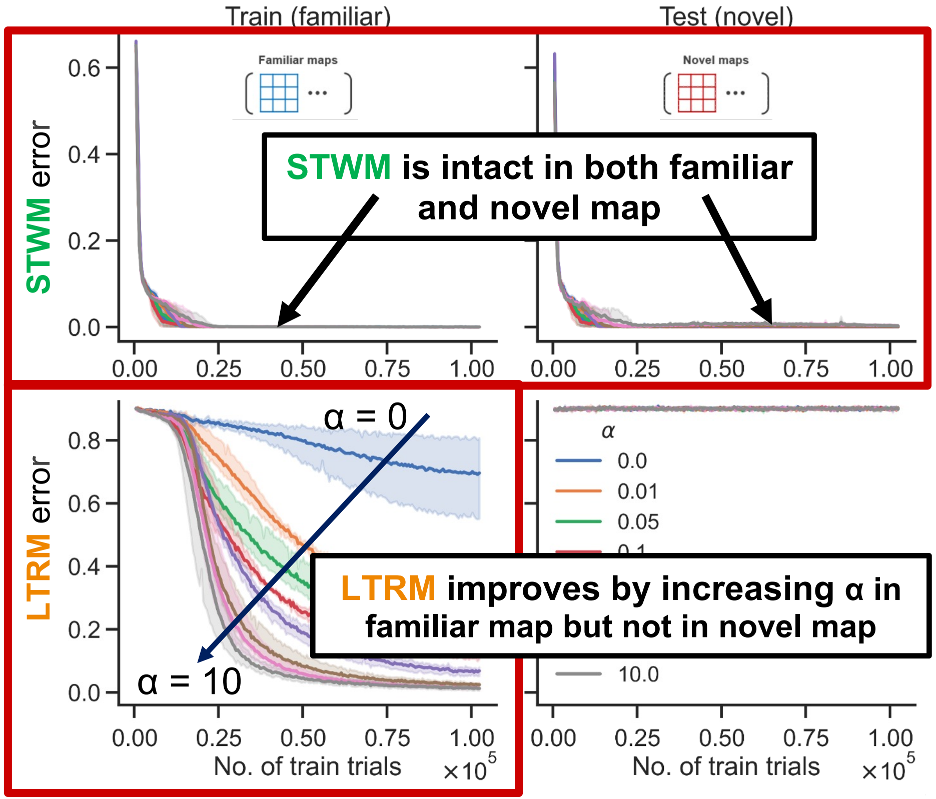
- We prepare *N* number of training maps.
- For each map, a random sensory observation among 10 letters are randomly placed at each node on each map.
- In a single trial, a randomly selected map from training maps is given to the agent.
- The agent starts at a random position and initiates a random walk on the selected map.

Two types of memory error in Transformer

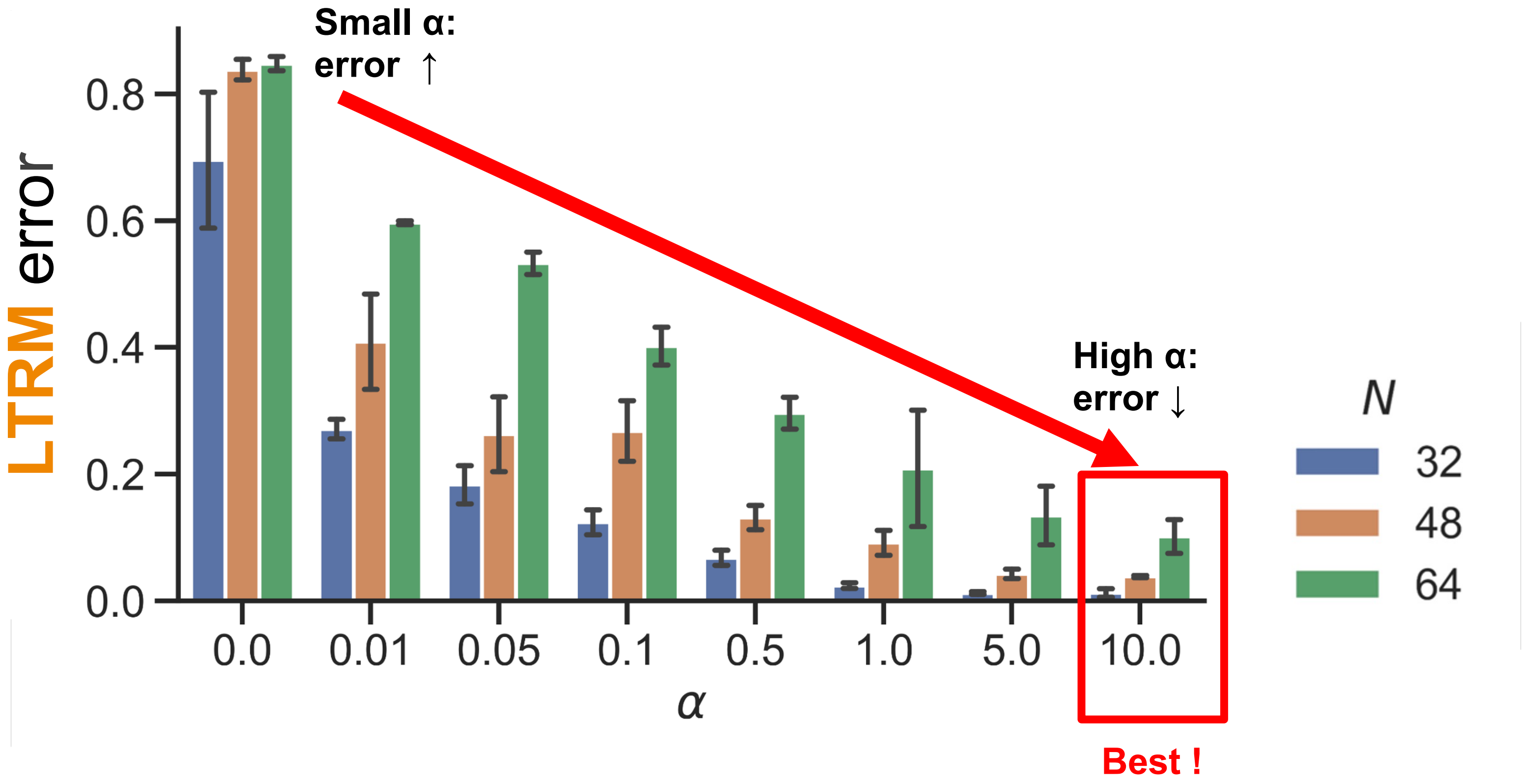


When answer is in the ***context*** - **STWM** error
When answer is **not** in the ***context*** - **LTRM** error

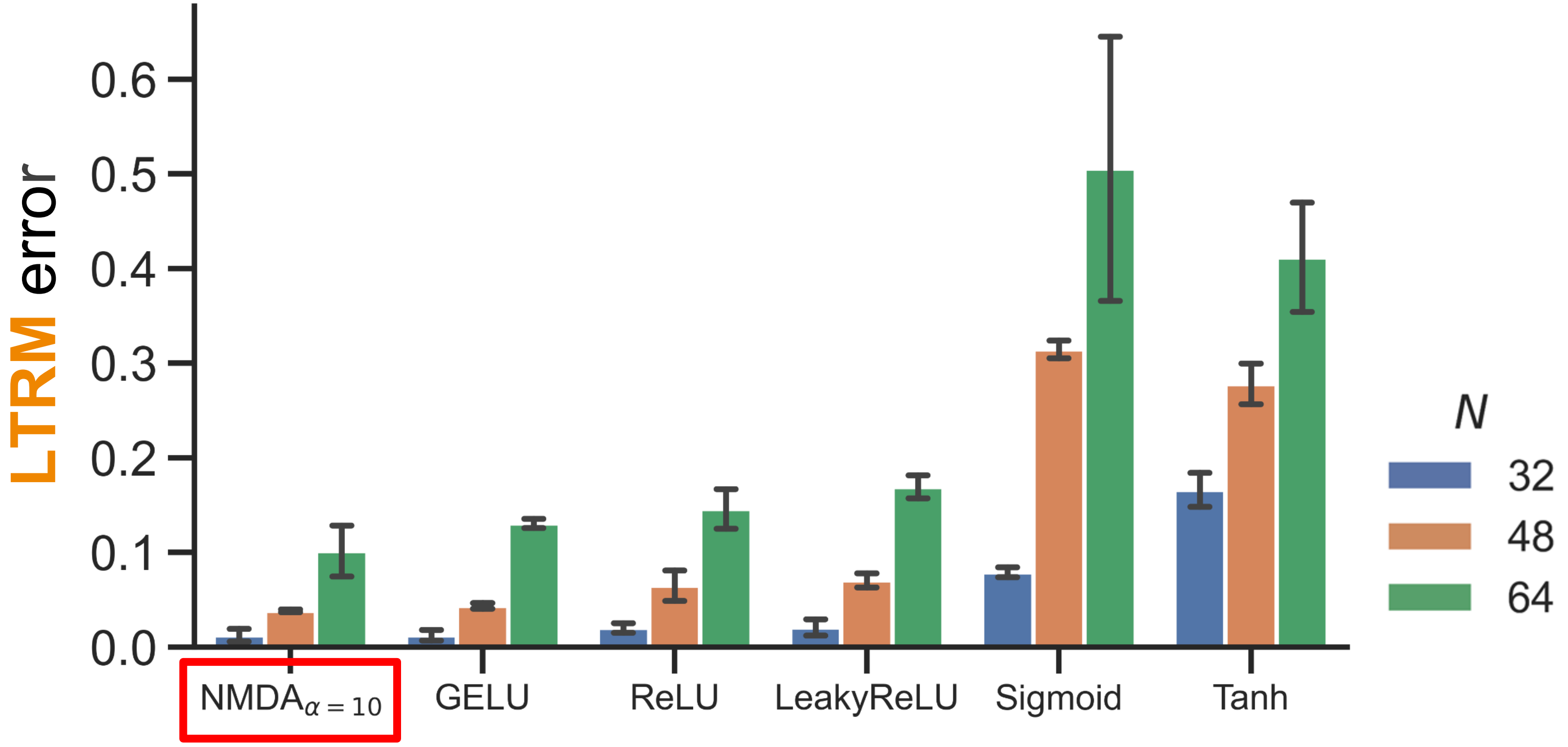
Our data show NMDA receptor nonlinearity controls long-term reference memory (LTRM)



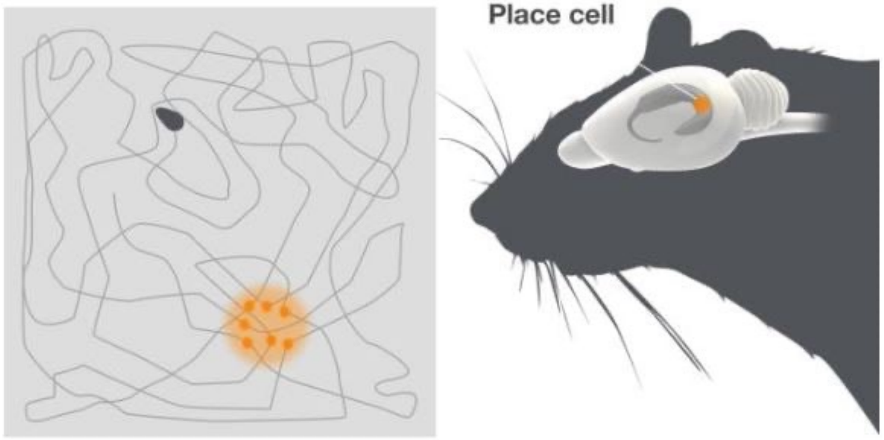
LTRM improves with increasing α (similar to how animal brain works)



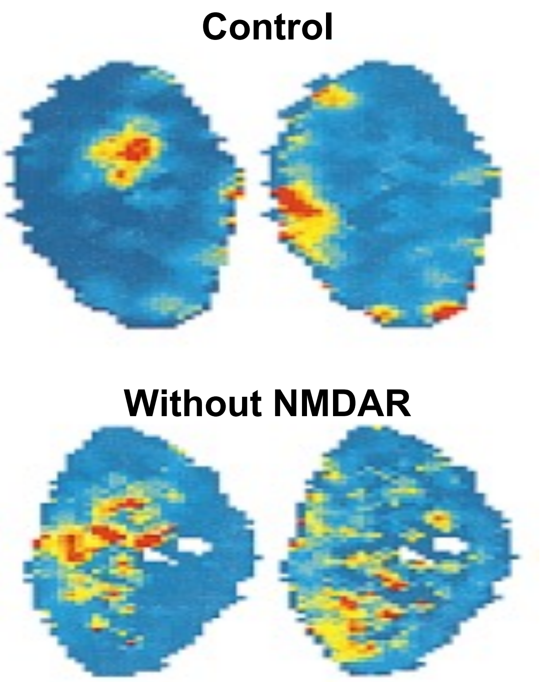
The improvement saturates after certain α (similar to how animal brain works)



RQ 2: Can NMDAR nonlinearity attribute to place cell emergence in Transformer?

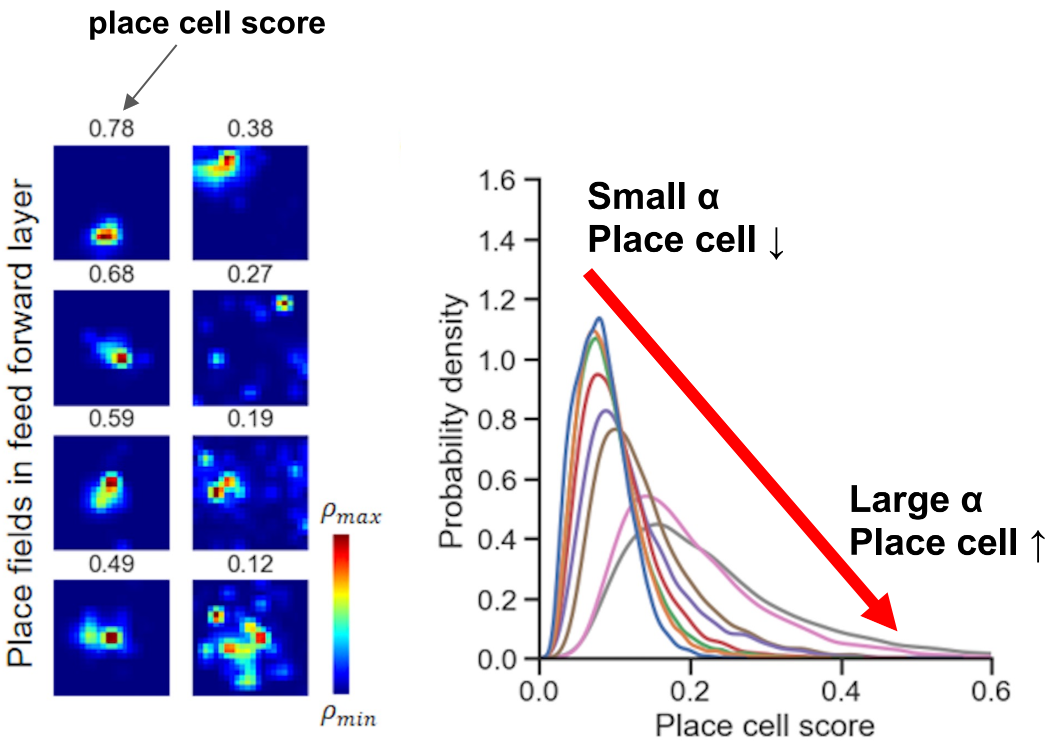


Creator: Ulrika Royen

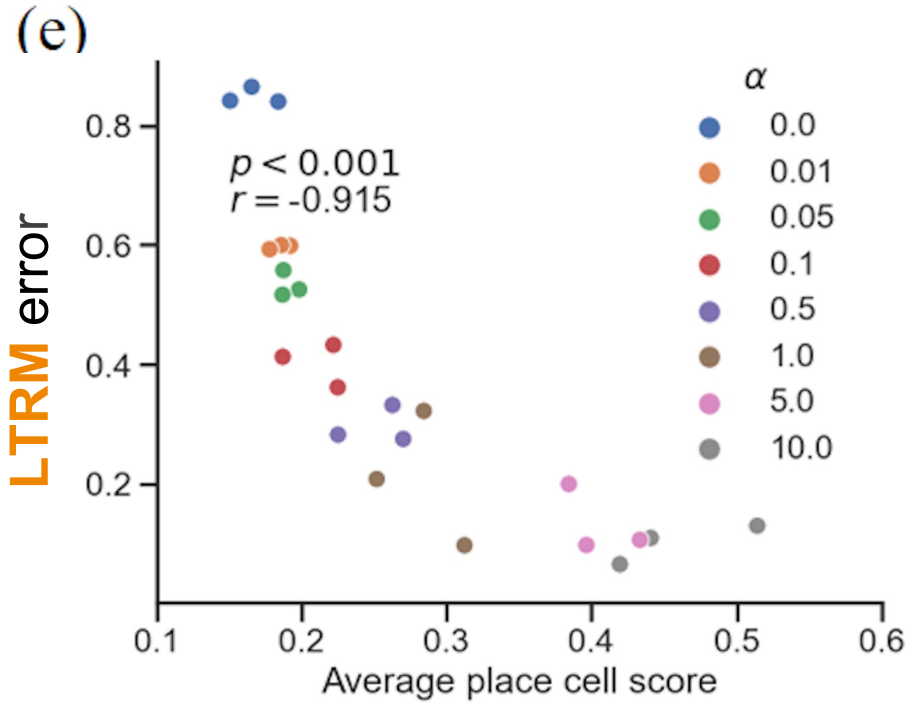


(McHugh et al., *Cell* 1996; Kentros et al., *Science* 1998)

Our data show place cells emerge in the Transformer’s FFN (feed-forward network)



Our data show transformers with high long-term memory forms more place cells (similar to brain)



Summary

1. We proposed a new activation function that is inspired by the nonlinear dynamics of NMDA receptors in brain (**NMDA**α) where α mimics the Mg^{2+} concentration level.

$$\mathrm{NMDA}_{\alpha}(x) = \frac{x}{1 + \alpha e^{-x}}$$

2. We developed a method for accessing the **reference memory**.
3. We evaluated the reference memory errors of **transformer** models with NMDAα. The results shows that *reference memory can be controlled by α*.
4. NMDAα with α=10 shows the **best reference memory performance** when compared to other widely used nonlinear activation functions.
5. We demonstrated the emergence of **place cells** in feed-forward networks of transformer for the first time.
6. Reference memory is impaired when the value of α in NMDAα is low and this **resembles long-term memory loss in brain**; low Mg^{2+} concentration in brain causes long-term memory loss.