Using a Biologically Plausible Long-Term Memory to Address Catastrophic Forgetting in AI

Bommer, S., Troquay, E. P. M., Zhang, X. & Starace, G. (2022)

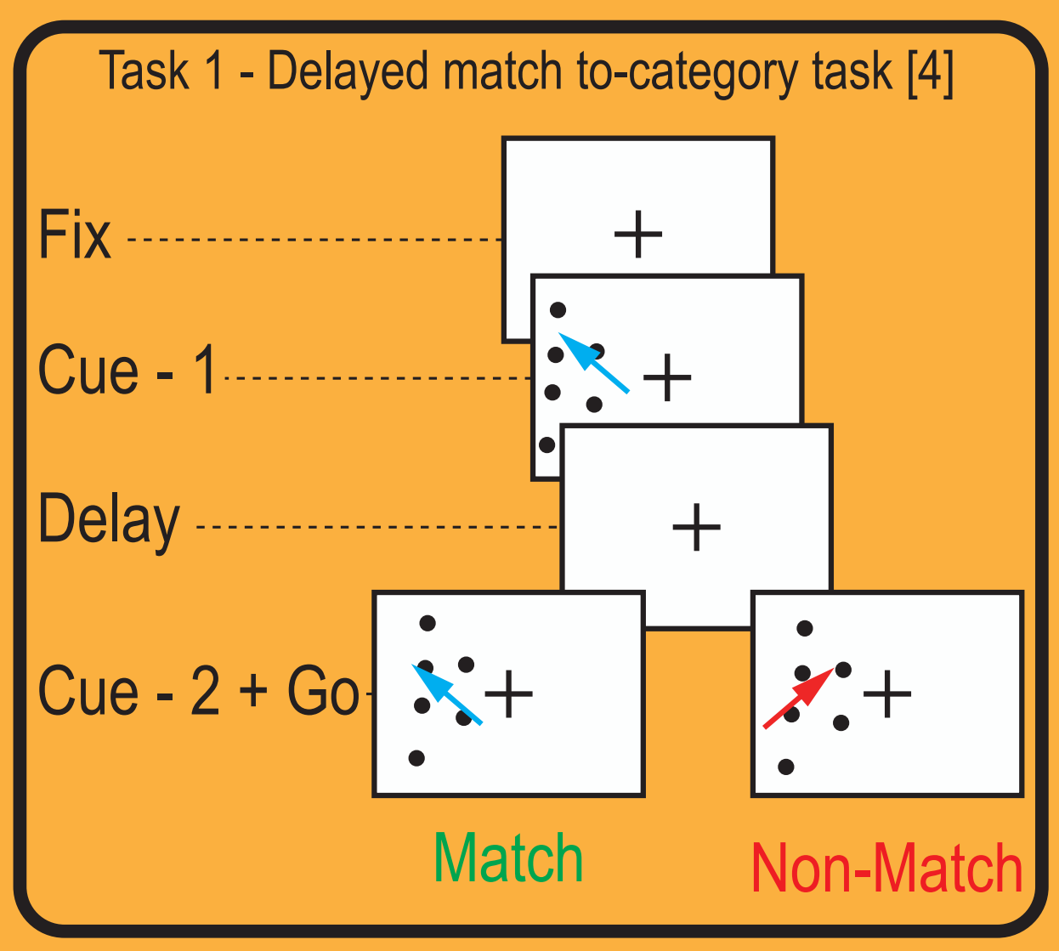
INTRODUCTION

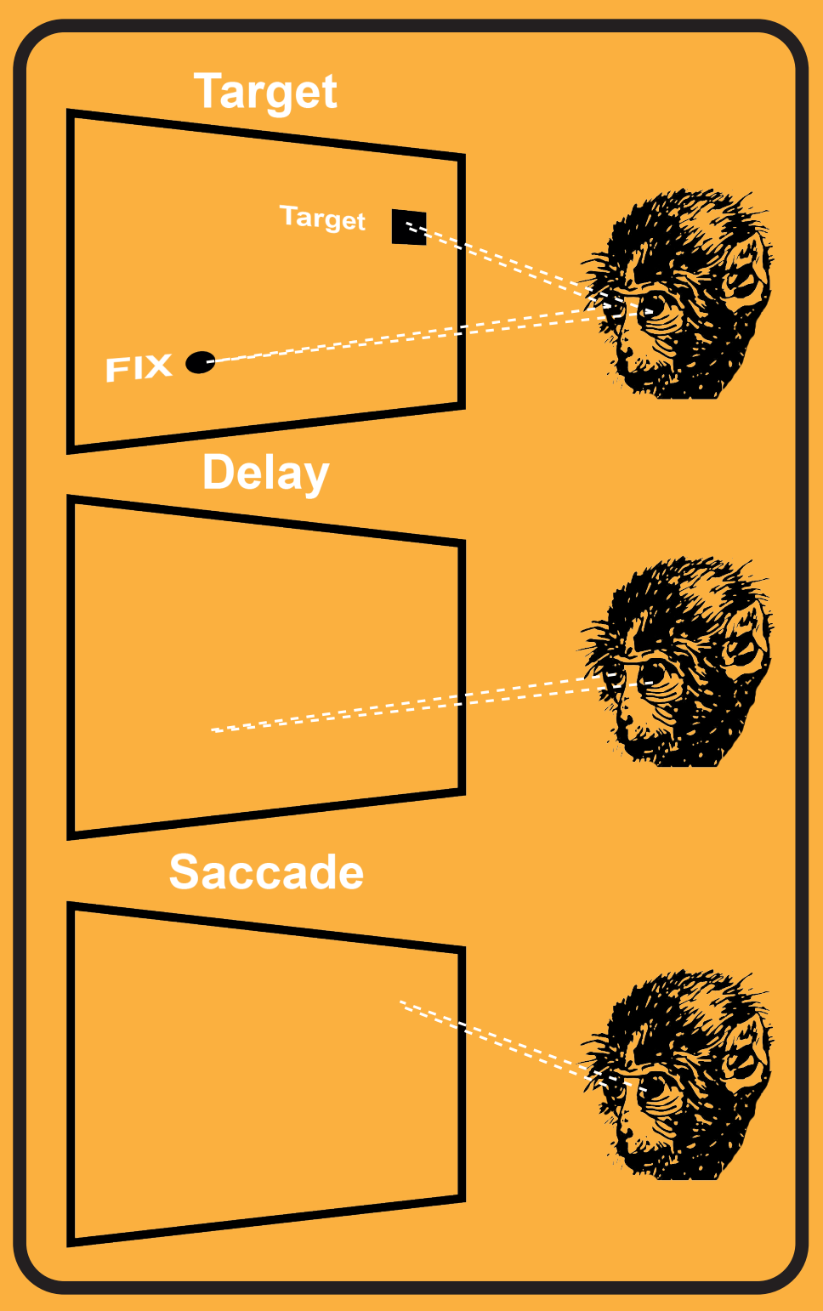
- ***Memory*** is an essential ingredient of general intelligence.
- The Atkinson–Shiffrin memory model [1]
  - Sensory register
  - ***Working memory (WM)***: temporary and limited store of information to be used in cognitive tasks that can be addressed with a small amount of steps
  - ***Long-term memory (LTM)***: a more permanent storage of information previously "rehearsed".
- ***Catastrophic forgetting*** is an issue commonly faced by computational models capable of working memory: ***performance on previously learned tasks is compromised upon learning a new task***.
- We tackle these questions from a ***biologically plausible*** perspective

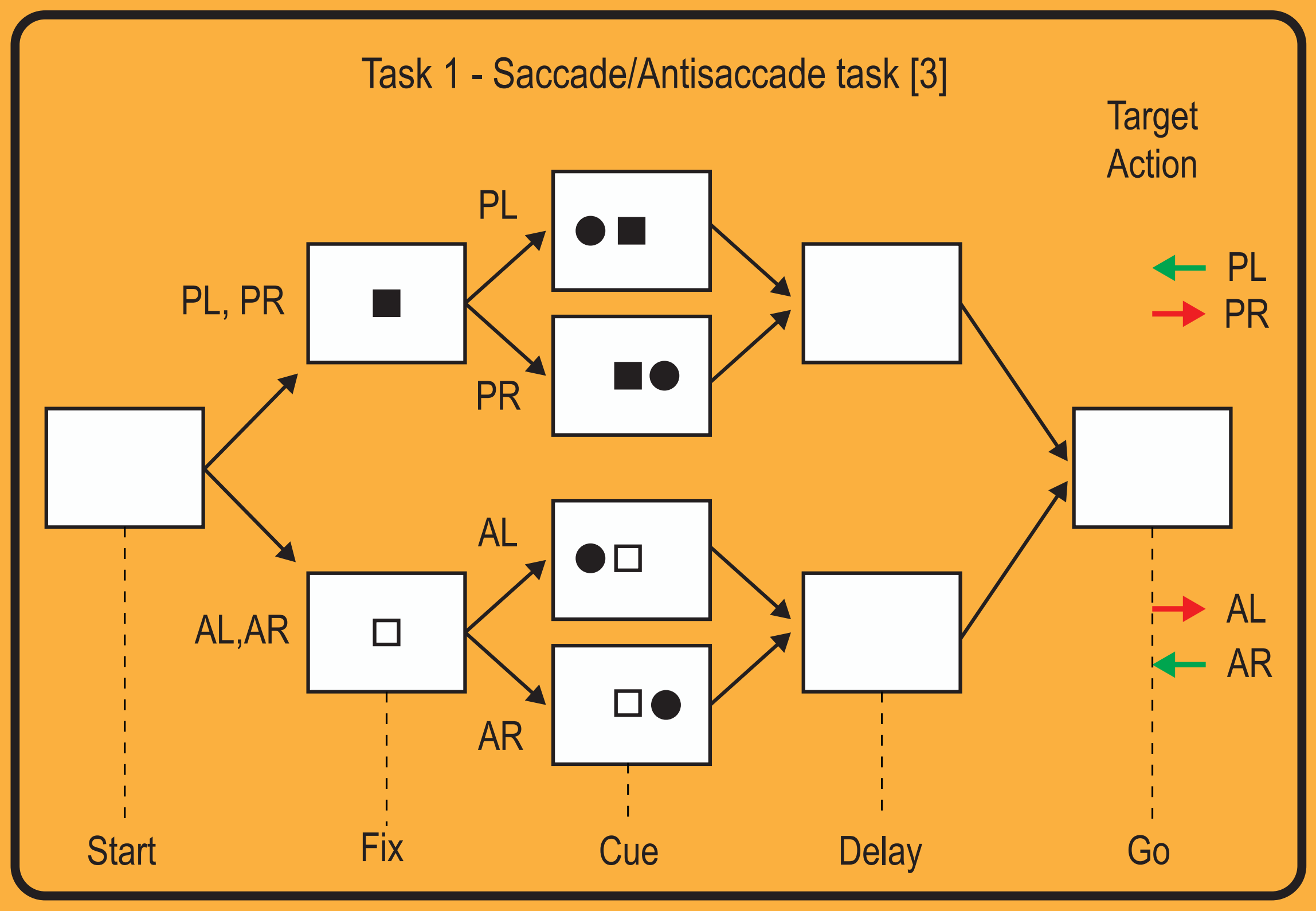
Research Question(s):

- ***Can Long-term memory address catastrophic forgetting?***
- ***Does equipping a computational model with Long-term memory more closely capture the memory mechanisms present in biological agents?***

Method

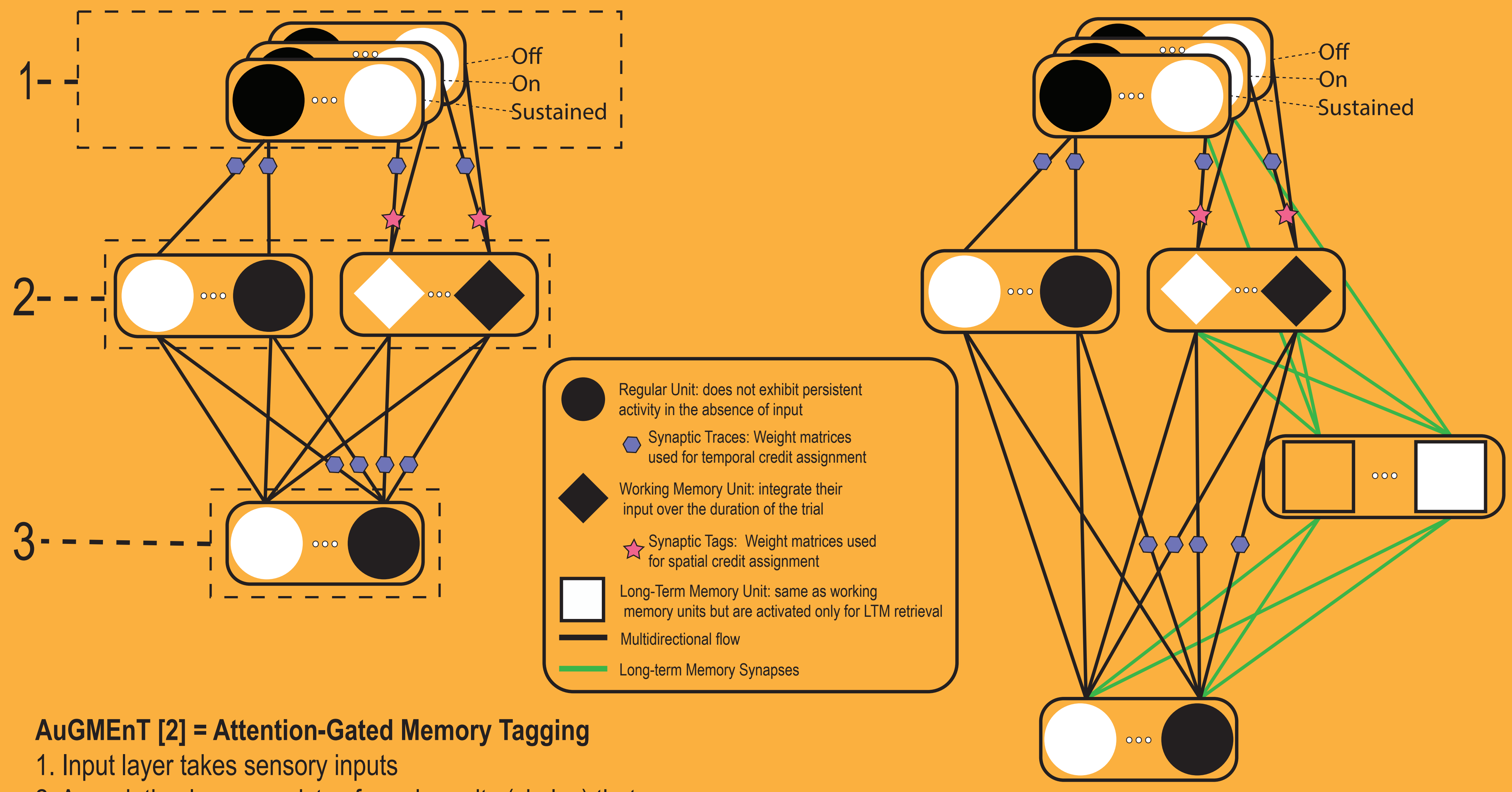






AuGMEnT

LTM-WM Model

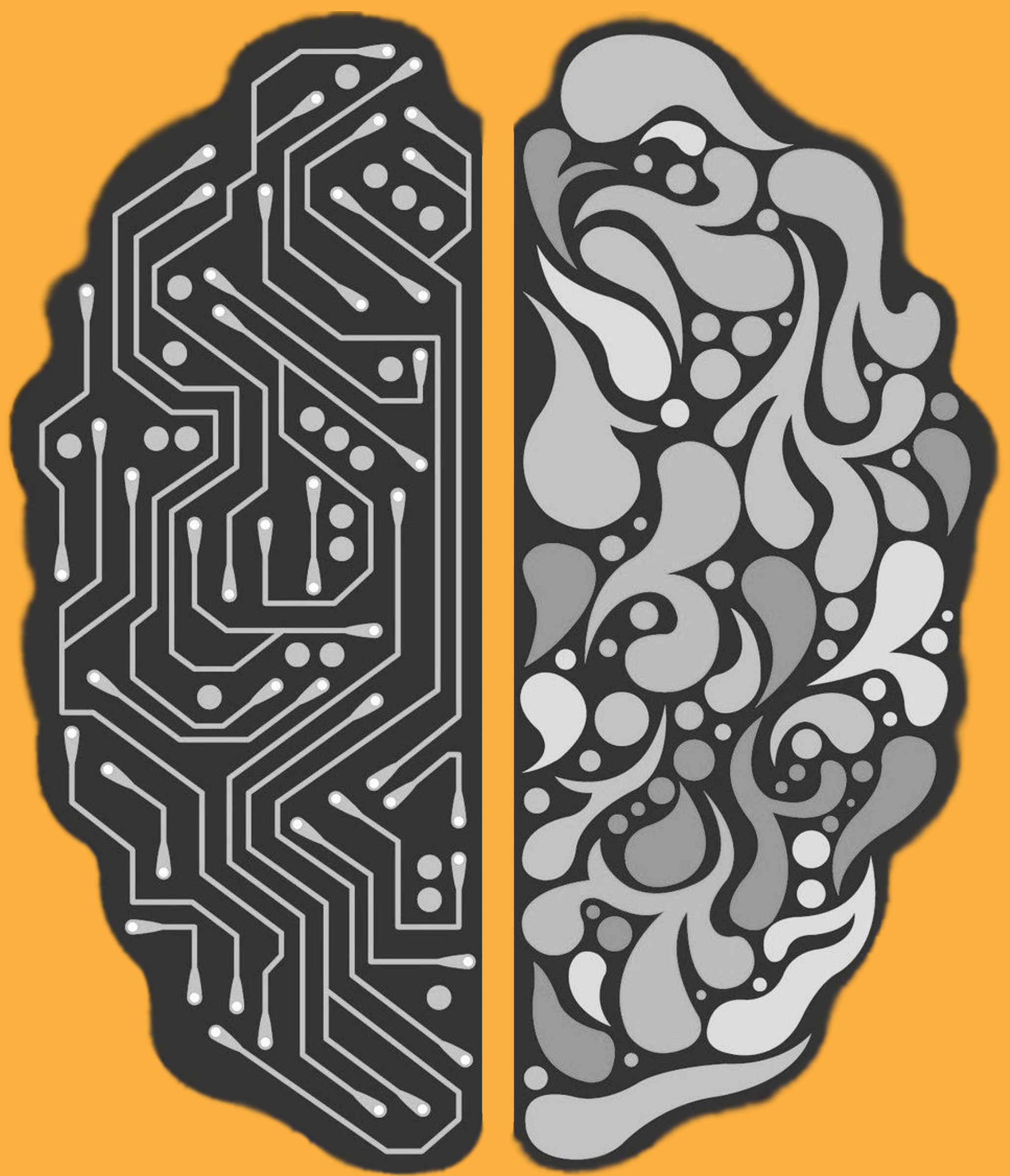

1
2
3
Off
On
Sustained
Off
On
Sustained
Regular Unit: does not exhibit persistent activity in the absence of input
Synaptic Traces: Weight matrices used for temporal credit assignment
Working Memory Unit: integrate their input over the duration of the trial
Synaptic Tags: Weight matrices used for spatial credit assignment
Long-Term Memory Unit: same as working memory units but are activated only for LTM retrieval
Multidirectional flow
Long-term Memory Synapses

AuGMEnT [2] = Attention-Gated Memory Tagging

1. Input layer takes sensory inputs
2. Association layer consists of regular units (circles) that process current sensory input, and WM units (diamonds) that integrate information over time
3. Output layer calculates Q values and decides action

LTM-WM = Long-Term Memory & Working Memory

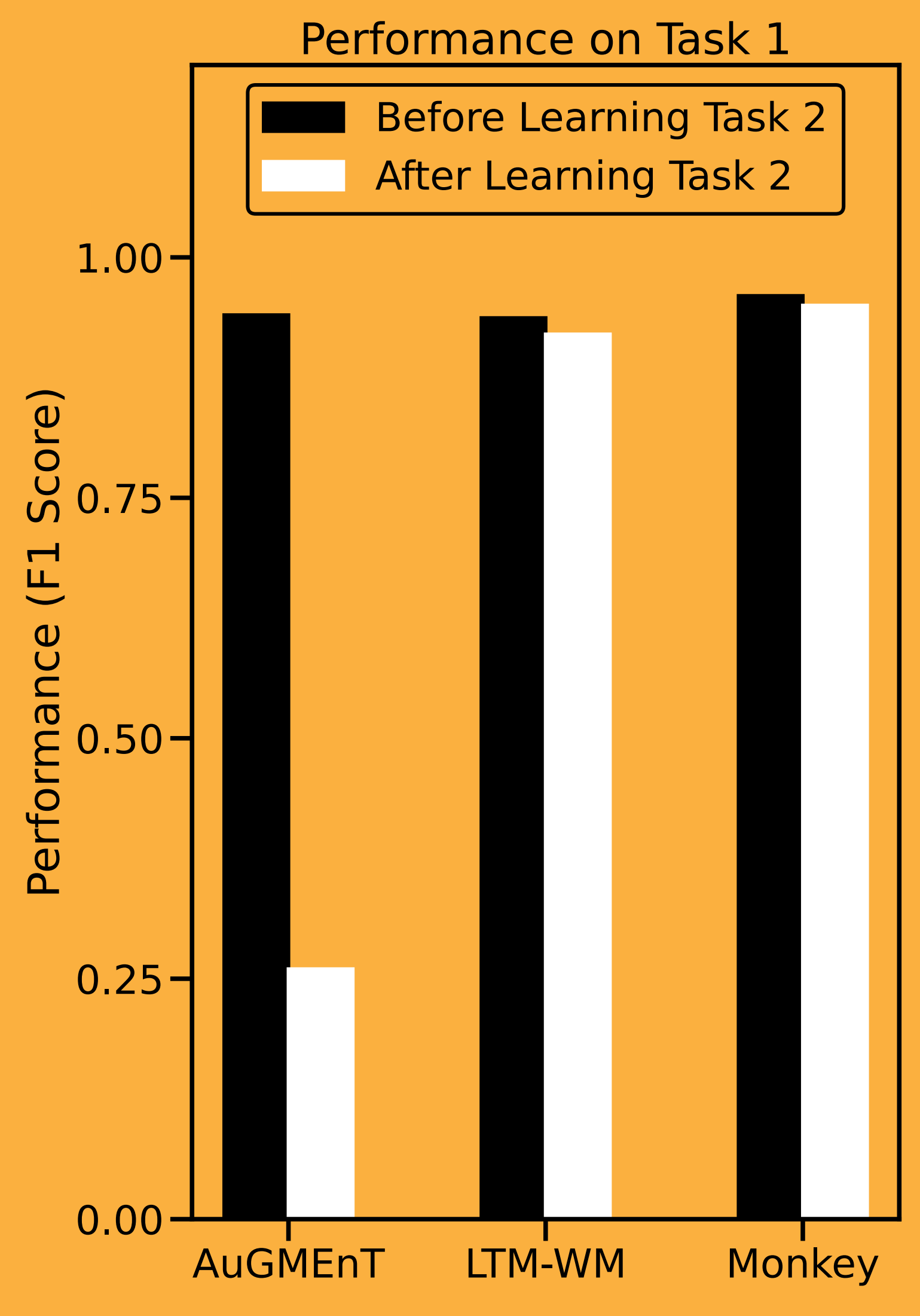
1. *Input layer* takes sensory inputs
2. *Association layer* consists of regular units that process current sensory input, *WM* units (diamonds) that integrate information overtime. Information stored in working memory units is subject to decay, but through *rehearsals*, it can be registered into *LTM units* (squares).
3. *Output layer* calculates Q values and decides action

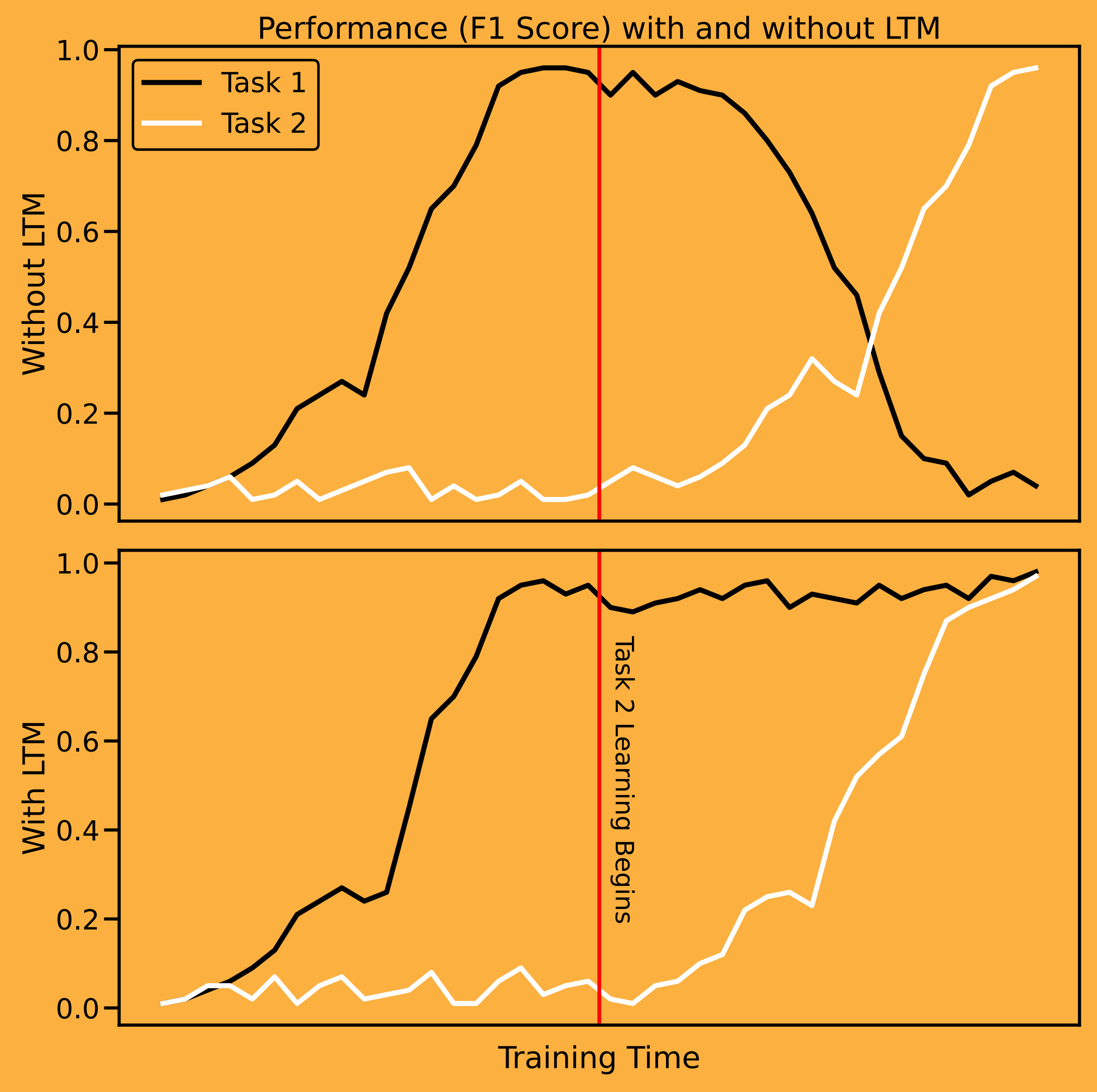


DISCUSSION

- As shown in the plot, after training for the second task:
  - AuGMEnT performance on task 1 *decreases*. This is because when AuGMEnT learns task 2, task 1 is *"forgotten"* as the weights are largely overwritten.
  - Meanwhile, the performance on task 1 is *retained* after learning task 2 for both the LTM-WM model and the monkeys. The model is able to recall the old task, as animals would do after successfully learning a task.
- In this sense, the LTM-WM model accounts for two critical properties of LTM [3], namely:
  - No temporal decay.
  - No chunk capacity limits.
- This suggests that LTM is essential to continual learning
- LTM-WM thus is a more biologically plausible representation of biological memory system, because:
  - When animals practice a task repetitively, the task is gradually consolidated [5] into their LTM, where retrieval becomes easier and more automatic. Even after switching to other tasks and back, monkeys are able to perform the old task.
  - Despite modeling the mechanism of WM, AuGMEnT still suffers from catastrophic forgetting. Once a new task is learned, the network "forgets" how to do the old task.
  - With LTM-WM, this problem is solved, hence the network becomes more biologically plausible.
- Future studies may focus on modeling other mechanisms involved in LTM (re)consolidation, like memory distribution over the cortex. Another example would be conditions that facilitate LTM consolidation, such as stress, or the role of es sential biological processes such as sleep [6].

EXPECTED RESULTS





MORE INFORMATION?

REFERENCES

[1] R. C. Atkinson and R. M. Shiffrin, 'Human Memory: A Proposed System and its Control Processes', vol. 2, K. W. Spence and J. T. Spence, Eds. Academic Press, 1968, pp. 89–195. doi: https://doi.org/10.1016/S0079-7421(08)60422-3.
[2] J. O. Rombouts, S. M. Bohte, and P. R. Roelfsema, 'How Attention Can Create Synaptic Tags for the Learning of Working Memories in Sequential Tasks', PLOS Computational Biology, vol. 11, no. 3, p. e1004060, Mar. 2015, doi: 10.1371/journal.pcbi.1004060.
[3] J. Gottlieb and M. E. Goldberg, 'Activity of neurons in the lateral intraparietal area of the monkey during an antisaccade task', Nat Neurosci, vol. 2, no. 10, pp. 906–912, Oct. 1999, doi: 10.1038/13209.
[4] D. J. Freedman and J. A. Assad, 'Experience-dependent representation of visual categories in parietal cortex', Nature, vol. 443, no. 7107, Art. no. 7107, Sep. 2006, doi: 10.1038/nature05078.
[5] N. Cowan, 'What are the differences between long-term, short-term, and working