COMPUTATIONAL MODELING OF PREFRONTAL-HIPPOCAMPAL INTERACTIONS IN VERBAL MNEMONIC DISCRIMINATION

by

Anuhya Suri

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
November 2023

To my father Subrahmanya Sai Suri, my source of inspiration, support and guidance.

Table of Contents

List of Tables

List of Figures

Abstract

Background: Mnemonic Discrimination, the ability to distinguish between highly similar memories is impaired in people with various neuropsychiatric disorders. While hippocampus and prefrontal cortex are attributed to perform this strategic process, there has been no study to understand the computational processes of executive control systems trained with reinforcement learning such as prefrontal cortex in this paradigm. In our study, we present a novel framework to study how prefrontal cortex lesions affect verbal mnemonic discrimination.

Methods: We employ a computational model to simulate a yes/no recognition task which is built upon Becker and Lim's [10] work on the verbal free recall model. We extend the model by representing words as continuous word embeddings to facilitate the calculation of mnemonic discrimination performance in verbal learning paradigms. We model hippocampus as a continuous Hopfield network and prefrontal cortex as a reinforcement learning agent for memory retrieval based on task requirements. We first validate our model by ensuring the free recall results are consistent with prior results. Using our novel implementation of a verbal mnemonic discrimination task, we then examine the role of prefrontal cortex lesions on mnemonic discrimination performance.

Results: Our results indicate that the lesions in prefrontal cortex have a significant negative impact on the mnemonic discrimination performance ($\beta = -0.07, p = 0.004$) and overall recognition performance ($\beta = -0.02, p = 0.014$).

Conclusion: Our study is the first one to build a computational model of interactions between the prefrontal cortex and hippocampus to study verbal mnemonic discrimination. Our study highlights the role of intact reinforcement based executive control system for mnemonic discrimination and recognition performance.

List of Abbreviations Used

CVLT	- California Verbal Learning Test
RAVLT	- Rey Auditory Verbal Learning Test
MST	- Mnemonic Similarity Task
MDI	- Mnemonic Discrimination Index
REC	- Recognition Index
FMRI	- Functional Magnetic Resonance Imaging
PET	- Positron Emission Tomography
DLPFC	- Dorsolateral Prefrontal Cortex
vmPFC	- Ventromedial Prefrontal Cortex
VLPFC	- Ventrolateral Prefrontal Cortex
mPFC	- Medial Prefrontal Cortex
RL	- Reinforcement Learning
TCM	- Temporal Context Model
MTL	- Medial Temporal Lobe module
PFC	- Prefrontal cortical module

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my research supervisor Dr. Abraham Nunes for giving me the opportunity to be a part of this research project. His constant guidance and support have been invaluable throughout this entire process. I am truly thankful for his prompt feedback on my work and his patience in providing me countless suggestions for improving my academic writing. I would like to extend my sincere thanks to my co-supervisor, Dr. Thomas Trappenberg for taking the time to provide me insightful suggestions and ideas to enrich my academic writing. I appreciate his critical feedback and mentoring which helped in shaping my thesis. I would also like to thank the members of my supervisory committee, Dr. Vlado Keselj and Dr. Carlos Hernandez Castillo for taking the time and effort in reviewing my thesis and providing me valuable feedback.

I am extremely thankful to my husband Karthik, without whom none of this would have been possible. His constant support, encouragement and motivation helped me to achieve this milestone and realize my dream. I am forever indebted to my parents, Subrahmanya Sai, and Kameswari, who instilled the importance of education in me and always encouraged me to pursue my dreams. Special thanks to my family and friends back home, who encouraged me and provided me with the emotional support I needed throughout this journey. Finally, I would like to thank my friends here at Dalhousie University who made this journey interesting and unforgettable.

Chapter 1

Introduction

Recognition memory is a vital cognitive function which gives us the ability to distinguish between old and new stimuli. A simple example of recognition memory is when we can recognize a person whom we have met in the past. Mnemonic discrimination is a specific part of the recognition memory mechanism which allows for the differentiation of new and previously encountered stimuli when they are highly similar. The hippocampus and the prefrontal cortex regions of the brain are attributed to perform these strategic processes. While various computational models have attributed the role of hippocampus to recognition memory and mnemonic discrimination [55, 13, 49, 50, 51], there is currently no computational model that explains the role of the prefrontal cortex, specifically in the verbal recognition paradigms.

The goal of this thesis is to build a biologically plausible computational model (refer Appendix F to understand the key motivation to build a computational model) that simulates the interactions between the brain's prefrontal cortex and hippocampus in order to predict how dysfunction in these interactions will impact mnemonic discrimination performance on widely-used clinical recognition memory tasks. This research will help us to better understand the mechanistic origins of neuropsychological deficits in clinical populations. From a computational perspective, this will also help us to understand how executive control systems trained with reinforcement learning, such as the human prefrontal cortex, affect the storage and recall capabilities of content-addressable memory systems [37], such as the human hippocampus.

To simulate the verbal recognition memory paradigm, we model the yes/no recognition task of the California Verbal Learning Test (CVLT) [1]. The CVLT is a widely used test in neuropsychological studies to investigate verbal recognition memory, as well as verbal recall, in humans. This means the focus of our study involves the use of words as stimuli in recognition where the model should predict if the word was encountered before or not.

We model the prefrontal-hippocampal interactions on the CVLT yes/no recognition task by extending an existing model of the CVLT free recall task by Becker and Lim [10]. We first extend the model to represent CVLT words as continuous representations using word embeddings. Additionally, we have implemented a continuous Hopfield network which can support this continuous input. We then include the CVLT yes/no recognition memory task to measure recognition memory performance and mnemonic discrimination. We propose three different approaches for the prefrontal-hippocampal interactions in the recognition memory task. The model consists of three layers (Fig. 3.1) each representing different regions of the brain that are involved in the recall and recognition process. The first layer is the lexical representation module where we model the CVLT words as continuous representation using ConceptNet word embeddings [68] as opposed to the binary representation of words in the Becker and Lim [10] model. The second layer is the medial temporal lobe module where we model the hippocampus as an autoassociative attractor network by using continuous Hopfield networks [58]. This gives the network two distinct capabilities: storage and retrieval [21]. The third layer is the prefrontal cortex layer which is modeled as a reinforcement learning agent that uses Q-learning algorithm to facilitate retrieval of memories based on task demands [10, 81].

Using this model, we first replicate the results of the CVLT free recall performance reported by Becker and Lim's [10] model. We compare our results with the Becker and Lim [10] model under three scenarios (1) intact model (2) lesioned model where nodes in the prefrontal cortex layer and its incoming and outgoing connections to medial temporal lobe layer are disabled and (3) blocked representation of input where words belonging to the same category are presented together. In each of these scenarios, we measure the recognition memory and mnemonic discrimination performance and compare the results across the three proposed approaches.

The results of the yes/no recognition task suggest that lesioning the prefrontal cortex has a statistically significant negative impact on the mnemonic discrimination as well as the overall recognition performance. This emphasizes the role of the intact prefrontal cortex in recognition memory paradigms. The mnemonic discrimination performance is enhanced due to full recollection (strategic retrieval of full representation of pattern including its context), but the overall recognition performance worsens indicating a trade-off between mnemonic discrimination and overall recognition performance. Further research is needed to address this trade-off between mnemonic discrimination and overall recognition performance. Nonetheless, our framework provides a solid foundation to study the role of the executive functions in the brain in recognition memory and mnemonic discrimination paradigms.

1.1 Research Question and Objectives

Our research focuses on building a computational model that studies how the prefrontal cortex and its interactions with the hippocampus influence the verbal recognition memory task. Since patients with neuropsychiatric disorders often exhibit prefrontal cortex lesions and experience verbal memory impairments [23], our study offers a valuable framework to analyze the impact of the prefrontal cortex lesions on verbal recognition memory.

Below is the research question, main objectives and contributions of this thesis.

1.1.1 Research Question

How do prefrontal cortical lesions impact verbal mnemonic discrimination performance in the CVLT recognition memory task?

1.1.2 Research Objectives

1. The primary objective is to propose different strategies for facilitating the prefrontal-hippocampal interactions in the CVLT yes/no recognition memory task

2. The secondary objective is to statistically evaluate the performance of the CVLT yes/no recognition task in recognition memory and mnemonic discrimination paradigms and to compare the performance among the three proposed approaches

3. The third objective is to assess the impact of prefrontal cortex lesions on the recognition memory and mnemonic discrimination performance and compare these results with the intact models

4. The fourth objective is to examine the influence of blocked presentation of the CVLT words on the yes/no recognition task and mnemonic discrimination for both the intact and lesioned models

1.1.3 Research Contributions

1. Developing a computational model that simulates the interactions between the prefrontal cortex and hippocampus in order to predict how the dysfunction in these interactions will impact the recognition memory and mnemonic discrimination paradigms
2. Employing word embeddings to encode the CVLT words as continuous representation, thereby enabling the simulation of recognition memory as applied to human trials
3. Introducing a method that extends the scope of CVLT's recognition memory paradigm allowing the evaluation of mnemonic discrimination within verbal recognition tasks

1.2 Thesis Outline

The remainder of this thesis is organized into 5 chapters.

1. Chapter 2 provides background for CVLT, measurement of mnemonic discrimination and neural underpinnings of recognition memory. It also includes previous research on the recognition memory in the brain and existing computational models which served as a foundation for our study.
2. Chapter 3 outlines the methodology designed for this study offering a detailed explanation of mathematical formulations used and a description of the experiments conducted.
3. Chapter 4 presents the results and conducts a comparative analysis of various experiments proposed in Chapter 3
4. Chapter 5 provides the discussion of the findings obtained in Chapter 4 providing the limitations of this study and proposing future directions for research.

Chapter 2

Background and Related Work

In this chapter we provide an overview of

1. How we can measure mnemonic discrimination through verbal recognition memory tasks
2. Overview of the California Verbal Learning Test
3. The neurological basis of recognition memory
4. Utilization of AI models to study recognition memory
5. Existing neuro-computational models for verbal learning tasks

2.1 Formalization of Mnemonic Discrimination in Verbal Recognition Memory Paradigms

In clinical assessments, the commonly used test for measuring mnemonic discrimination ability is the Mnemonic Similarity Task (MST) [69] (refer Appendix A for more information), which is an object recognition task using image stimuli. Since verbal memory is crucial for daily human functioning [22, 75, 16], it is important to study the mnemonic discrimination ability in the context of verbal learning (refer Appendix B for more information). Moreover, many commonly used neuropsychological tests that assess memory, like CVLT and Rey Auditory Verbal Learning Test (RAVLT) [9] focus on verbal learning paradigms. Consequently, there are abundant clinical data available for in depth analysis in this area [33]. In this section we aim to provide an intuition on the formalization of the mnemonic discrimination in verbal recognition memory paradigms.

In order to create a computational model of human behavior on recognition memory tasks, we must mathematically formalize these paradigms such that a suitable testing environment can be modeled. Furthermore, to understand the computational nature of mnemonic discrimination and how it should be measured, we must derive such a measure from a suitable formalism of recognition memory tasks in general. We follow the intuition presented in Leger et al. [40] to formalize recognition memory tasks, and from this formalism derive an understanding of, and measurement index for, mnemonic discrimination.

A verbal recognition memory experiment consists of an encoding phase (also known as study phase) and a test phase. In the encoding phase, the agent is supplied with a sequence of N items which it must memorize. We denote the list of N items presented during the encoding phase as $Y_s = (y_{s_1}, y_{s_2}, ..., y_{s_N})$. In the test phase, the agent is presented with recognition test items, one at a time, which it must correctly classify as *old* or *new.* Let $Y_r = (y_{r_1}, y_{r_2}, ..., y_{r_M})$ be the sequence of M items in the recognition test list. Each item in Y_r can either be one of the items in Y_s or not.

Given the lists Y_s and Y_r, let $X^* = (x^*_1, x^*_2, ..., x^*_M)$ be a binary vector denoting

$$x^*_i = \mathbb{1}[y_{r_i} \notin Y_s] \tag{2.1}$$

which takes a value of 1 if the recognition item is a novel stimulus that was not memorized as part of the study list and 0 if the item was previously encountered. Given the ground truth x^*_i, let x_i be the agent's prediction of whether the i^{th} recognition item is part of Y_s or not. Let us consider

$$x_i = f_A(y_i; Y_s) \tag{2.2}$$

where $f_A(y_i; Y_s)$ performs recognition memory judgements and generates predictions.

Let $d(y, Y_s)$ be the distance between a test list item y and the perceptually or semantically *closest* stimulus encoded list Y_s such that $d(y, Y_s) = 0$ indicates that y is a target (part of the encoded list). This implies that y is by definition a novel stimulus if $d(y, Y_s) > 0$. However, as $d(y, Y_s)$ increases, the degree of perceptual or semantic *novelty* increases. That is, the novel stimulus y becomes more different from the stimuli previously encountered in the study list. We measure the mnemonic discrimination capability by considering how the probability of identifying an item y as novel, varies with the degree of similarity between y and the encoded list Y_s. If we let $p_A(y)$ denote the probability that agent A classifies stimulus y as novel, then we assume that

$$p_A(y) \propto d(y, Y_s) \tag{2.3}$$

This means that the more the recognition item y is distant from the encoded item, the more likely it is to be classified as *new*. This assumption is well substantiated in the literature [69]. For agents with perfect mnemonic discrimination capability, $p_A(y)$ will be maximum even for small values of $d(y, Y_s)$ indicating that the recognition memory is sensitive to small changes in the stimuli.

Having formalized the mnemonic discrimination in the context of verbal recognition tasks, in the next section we will introduce the California Verbal Learning Test (CVLT). We will elaborate on how recognition performance and mnemonic discrimination are measured in CVLT specifically in the context of our computational model.

2.2 The California Verbal Learning Test (CVLT)

2.2.1 Phases of the CVLT

The California Verbal Learning Test (CVLT) is among the world's most commonly used neuropsychological tasks in clinical settings [41, 7]. The CVLT test consists of 8 subtasks involving of a series of immediate and delayed recall and recognition tests, which are listed here, but described in further detail below:

1. List A Immediate Free Recall

2. List B Immediate Free Recall

3. List A Short-Delay Free Recall

4. List A Cued Recall

5. List A Long-Delay Free Recall

6. List A Long-Delay Cued Recall

7. List A Yes/No Recognition

8. List A Forced Choice Recognition

List A Immediate Free Recall (IFR-A) consists of an encoding phase (also called study phase) and recall phase which happens over a course of $N_{trials} = 5$ trials. During the encoding phase, the agent is presented with 16 List A words, which are organized into 4 categories: vehicles **(truck, motorcycle, subway, boat)**, vegetables (**spinach, onion, celery, cabbage**), animals **(giraffe, zebra, cow, squirrel)**, furniture **(bookcase, cabinet, lamp, desk)**. However, despite the organization into four semantic categories, the words are read in a fixed, shuffled order: **(truck, spinach, giraffe, bookcase, onion, motorcycle, cabinet, zebra, subway, lamp, celery, cow, desk, boat, squirrel, cabbage)**. Immediately after being read the 16 words, the agent must recall the words from memory. Recall can proceed in any order. We denote the number of attempts made to recall words at trial t as $N_{attempts}$. Each recall phase is terminated when the agent cannot remember any more words and so the number of attempts in each recall trial varies.

List B Immediate Free Recall (IFR-B) presents a new list of 16 words to the agent, which it must then recall immediately. It consists of an encoding phase (also called study phase) and recall phase which happens over a single trial. During the encoding phase, the agent is presented with 16 words sequentially, which are split into 4 categories musical instruments **(violin, guitar, clarinet, saxophone)**, locations in homes **(closet, basement, garage, patio)**, vegetables **(cucumber, turnip, corn, radishes)**, animals **(elephant, sheep, rabbit, tiger)**, such that there are 2 categories that overlap with List A, and 2 categories that are distinct. The words are presented in a fixed, shuffled order: **(violin, cucumber, elephant, closet, turnip, guitar, basement, sheep, clarinet, garage, corn, rabbit, patio, saxophone, tiger, radishes)**. During the recall phase, the agent tries to reproduce, from memory, the words learned during the encoding phase. The recall trial ends when the agent cannot remember any more words. The purpose of IFR B is to serve as an interference task for the rest of the tasks occurring later in the CVLT test.

List A Short-Delay Free Recall (SDFR-A) occurs immediately after IFR-B, the agent is once again asked to recall List A, based only on the first five encoding trials done during IFR-A.

List A Short-Delay Cued Recall (SDCR-A) involves asking the agent to recall List A in a cued fashion. That is, the agent is asked to recall all words that were furniture, vegetables, vehicles, and animals.

List A Long-Delay Free Recall (LDFR-A) occurs following SDCR-A and a 20 minute delay. LDFR-A involves once again recalling as many List A words as possible learned during the IFR-A phase.

List A Long-Delay Cued Recall (LDCR-A) occurs following LDFR-A, under the same category prompts as SDCR-A.

List A Yes/No Recognition (YNR-A) occurs following LDCR-A, and involves presenting the agent with 48 words, which it must identify as having been on List A (*Yes*) or not (*No*). These words are divided into four sets:

1. true targets (those that were on List A)
2. those from List B with categories that are shared with List A
3. those from List B with categories not shared with List A
4. totally novel words

List A Forced Choice Recognition (FCR-A) occurs immediately after YNR-A. Here, agents are presented with two words and asked which was on List A. Since most subjects in psychiatric samples obtain perfect scores on forced choice recognition, we will not model it in the present study [7, 15].

2.3 Performance Evaluation on the CVLT

2.3.1 Free Recall Performance Evaluation

For some recall attempt in any of the free or cued recall phases above (IFR-A, IFR-B, SDFR-A, SDCR-A, LDFR-A, LDCR-A), we define $q = (q_i)_{i=1,2,\ldots,N_{attempts}}$ as a boolean vector, where $q_i = 1$ if the i^{th} recalled word was present in the study list and is not a repetition (an instance where a given word has been recalled before) or intrusion (an instance where a given word was not part of the study list). The **total number of correct recalled words** $\hat{C}$ in each trial is calculated by

$$\hat{C} = \sum_{i=1}^{N_{attempts}} q_i \tag{2.4}$$

Since IFR A task has 5 trials, the total correct recalled words from trial 1 to trial 5, denoted by C is calculated as:

$$C = \sum_{i=1}^{N_{trials}} \hat{C}_i \tag{2.5}$$

This statistic is often used as a global summary of verbal memory performance in clinical samples, and so forms the primary outcome measure for free-recall components of our study.

Another important measure of performance on the CVLT which is reflective of good executive control is **semantic clustering**, which identifies the degree to which recalled words *cluster* into their semantic categories during free recall.

The Semantic Clustering (Observed) score measures the degree to which correct words from the same category are recalled in close temporal proximity [1]. To calculate the total semantic clustering score per trial, we need to sum the total number of correct responses satisfying the above condition. Let $sc = (sc_i)_{i=1,2,...N_{attempts}}$ be a boolean vector representing if the recalled word belongs to the same category as the previous recalled word in a given trial. By default, the first element sc_1 will be 0 since it is the first word recalled. The **semantic scores (observed)** denoted by OS, is calculated as

$$OS = \sum_{i=1}^{N_{attempts}} \mathbb{1}(q_i = q_{i-1} = sc_i = 1) \tag{2.6}$$

2.3.2 Mnemonic Discrimination and Recognition Performance Evaluation in the Delayed Yes/No Recognition Phase

We assess the mnemonic discrimination and overall recognition performance in the List A Yes/No Recognition (YNR-A) phase. To assess the mnemonic discrimination performance on the CVLT, we adopt the mnemonic discrimination measure proposed by Leger et al. [40] that is designed specifically for tests such as CVLT that do not specify a categorical distinction between *lures* (words that are semantically related to List A words but are not part of it) and *foils* (novel words). In this case mnemonic discrimination performance can be measured as a function of similarity between recognition test words and the encoded words which is List A words for the CVLT test.

Following the formalization of the mnemonic discrimination paradigms in Section 2.1, Eq. (2.3) indicates that the probability of classifying a stimulus as *new* increases as distance of the word increases from the encoded list (List A words of CVLT). The distances $d(y, Y_s)$ are scaled between 0 and 1.

The mnemonic discrimination ability is specifically assessed for test words that are highly similar to List A words, that means their distance from list A words is minimum. In the CVLT recognition test, one such word is *carrot* which is not part of list A words but is highly similar to vegetables in the List A words like spinach, onion, celery, cabbage. Any subject with high mnemonic discrimination ability can distinguish words like carrot as *new* making them highly sensitive to small differences in stimuli.

Therefore, we need to model the probability of classification as a function of similarity with the encoded (ListA) words. Let P_{new} be the probability that a recognition word is classified as *new*. The influence of similarity (distance) on the probability P_{new} is modeled as follows by Leger et al. [40]:

$$P_{new}(Distance) = d + (a - d)/(1 + Distance/c)^b)^e \tag{2.7}$$

The parameter a is the lower asymptote of the curve which influences the probability that an *old* word is misclassified as *new*. The parameter b represents the slope of the curve which indicates how sharply the words are distinguished compared to previous words. The parameter c is the horizontal shift in the curve. The parameter d is the upper asymptote of the curve which influences the probability that a *new* word (distant from List A words) is correctly classified as *new*. The parameter e allows flexibility in the curve by introducing asymmetry of the sigmoidal function.

Based on this, Leger et al. [40] calculates the Mnemonic Discrimination Index (MDI).

$$MDI = 1 - \frac{A}{REC} \tag{2.8}$$

where

$$REC = P_{new}(1) - P_{new}(0) \tag{2.9}$$

and

$$A = P_{new}(1) - \int_0^1 P_{new}(x)dx \tag{2.10}$$

We use the MDI to calculate the mnemonic discrimination performance and the REC to get the overall recognition performance. Findings from Leger et al. [40] demonstrate that MDI and REC strongly associate with the measurements in the gold standard MST. Therefore, these provide valid and well-seperated measurements for mnemonic discrimination and overall recognition performance respectively.

The next step involves defining the function $f_A(y_i; Y_s)$ in Eq. (2.2) required to execute the recognition memory judgements. To achieve this, it is important to gain insight into the underlying processes of recognition memory in the brain. This will serve as a foundation to develop a computational model capable of performing this function in a way that may help us to understand the neural basis of recognition memory in humans. In the next section, we describe the neural mechanisms of recognition memory in the brain.

2.4 Neurological Basis of Recognition Memory and Mnemonic Discrimination:

The process of making recognition memory judgements, which involves distinguishing between *old* and *new* stimuli is a multi-stage process. Various regions of the brain have distinct functions in facilitating these computational processes. We propose a computational framework which involves two important regions of the brain hippocampus and prefrontal cortex that interact together to make successful recognition memory judgements. Furthermore, we also look at how the interaction between these two regions lead to a successful mnemonic discrimination capability. Below is a detailed description of the role of hippocampus and prefrontal cortex in the recognition memory. We also present previous studies that provide evidence for the role of these regions in recognition memory and mnemonic discrimination.

2.4.1 Role of Hippocampus

When the brain is presented with any stimulus and is tasked with distinguishing whether it is *old* or *new*, it initiates a recollection process [43, 85, 32] which involves strategic search in memory to get the contextual details of the stimulus. Successful *recollection* depends on proper encoding of memories and retrieving context appropriate memories. For example, in the CVLT, when the participant listens to the words, successful encoding takes place in the brain. The yes/no recognition task requires the brain to retrieve the encoded pattern that corresponds to the cue presented to the participant. If the brain successfully retrieves the pattern, the participant responds *old* or else responds *new*. These two stages of encoding and retrieval require a computational process called *pattern completion* that takes place in the hippocampus. Pattern completion allows for accurate generalization of any pattern when presented with a noisy or partial cue [21].

The network architecture of the CA3 region of the hippocampus facilitates the pattern completion processes. The recurrent interconnected pyramidal cells of the CA3 [6] operate as an autoassociative network enabling storage of patterns and retrieval if a partial cue is presented to it [21, 60]. When a partial pattern is presented, subsets of the CA3 neurons are activated subsequently retrieving the whole pattern thus resulting in pattern completion.

Numerous studies have emphasized the role of hippocampus in recognition memory and mnemonic discrimination performance. Researchers frequently investigate these aspects through Functional Magnetic Resonance Imaging (FMRI) studies, which examine the activity of the hippocampus during recognition tasks, or by evaluating the results of recognition tasks when individuals with hippocampal lesions engage in the recognition tasks. In the next section we will provide an overview of past studies that have explored the role of hippocampus in recognition memory and mnemonic discrimination performance.

Previous Studies on the Role of Hippocampus in Recognition Memory and Mnemonic Discrimination

FMRI studies in the human brain conducted by Bakker et al. [6] and Klippenstein et al. [36] has provided evidence that the CA3 region of the hippocampus is activated when lures or new items are presented during the recognition task. This strongly supports the involvement of CA3 in storing distinctive representations of patterns which is required for successful discrimination of stimuli.

Bayley et al. [8] investigated memory impaired patients with hippocampal damage by conducting a yes/no recognition task focused on object recognition. In this task, the participants initially should identify if the presented images are living/man-made followed by a yes/no recognition task and a forced-choice recognition task. The results from the yes/no recognition task indicated that patients with hippocampal damage performed poorly when the test items are highly similar to the study items impacting mnemonic discrimination performance.

Yassa et al. [84] conducted a FMRI study to examine the performance of memory-impaired older adults in an object recognition task where they had to classify images as *old*, *new* or *similar*. The findings revealed that the CA3 region in the older adults showed reduced activity when dealing with lure items that are highly similar to target items as compared to young adults, whose CA3 region responded more robustly even when the lure items are highly similar to the targets. These results suggest that the age related microstructural changes in the CA3 region can lead to behavioral discrimination deficits in mnemonic discrimination tasks.

Manns et al. [44] conducted a verbal recognition memory task using the verbal learning test RAVLT on seven patients with bilateral hippocampal damage. All patients obtained poorer recognition scores on the RAVLT test. In the next experiment participants were asked to not only determine if they had encountered an item previously but also to indicate if their judgment was based on recollection or simple familiarity which does not involve any strategic search in memory. The results showed that all participants were impaired similarly in both recollection and familiarity, suggesting the essential role of intact hippocampus for recognition memory tasks that requires both recollection and familiarity aspects.

2.4.2 Role of Prefrontal Cortex

The prefrontal cortex is characterized as the region responsible for executing strategic functions based on specific requirements of a given task [48]. For example, when the recollection involves semantic organization, serial organization of words, or cue based recollection, the prefrontal cortex biases the medial temporal lobe's retrieval to facilitate organized recollection [57, 53].

Previous Studies on the Role of Prefrontal Cortex as a Reinforcement Learning Agent

Frith [20] defines the role of Dorsolateral Prefrontal Cortex (DLPFC) as selecting an appropriate response when there are multiple alternate responses, as opposed to situations where only one response is viable. This is based on the Positron Emission Tomography (PET) study on DLPFC conducted by Nathaniel and Frith [52] where participants engaged in a word generation test consisting of two tasks (1) generating a word that should fit a sentence and (2) generating a word that should not fit a sentence. Each task again had sentences with varying degrees of constraint. For example, *He mailed the letter without a....* is a highly constrained sentence because 98% of the participants complete this sentence with the word *stamp*. A sentence like *The police had never seen a man so....* has minimal constraint because it can have many alternate responses (e.g. nervous, violent, upset) [52]. When the DLPFC was monitored during these tasks, it showed higher levels of activation during tasks with low constraint and in situations where the answer should not fit the sentence (task 2). In both these scenarios where answers were not immediately evident and required participants to choose from numerous words known to them, Frith referred to this response selection as *sculpting the response spac*e which captures the main role of DLPFC.

Duncan [17] defines the fundamental role of the prefrontal cortex to be adaptive neural coding where the neural representations in the prefrontal cortex can dynamically adapt to suit the specific demands of the task at hand. This adaptability enables the prefrontal cortex to emphasize relevant inputs and filter out irrelevant responses, especially in the presence of many alternate responses. Duncan further suggests that the relevance of information in any task in the prefrontal cortex is determined based on rewards.

The process of reward-based learning in the prefrontal cortex is attributed to the phasic signals provided by dopamine that conveys the reward-prediction error influencing actions and learning in the prefrontal cortex [54, 81, 73]. Based on these views, the prefrontal cortex has the capability to develop mnemonic codes rapidly based on task demands and self-monitor its performance based on current state and goals [10].

In the following section, we will explore previous studies that have examined the role of prefrontal cortex and its various sub-regions in recognition memory paradigms.

Previous Studies on the Role of Prefrontal Cortex in Recognition Memory and Mnemonic Discrimination

Lauzon [39] specifically studied the impact of lesions in the Ventromedial Prefrontal Cortex (vmPFC) on mnemonic discrimination performance. 10 adults with vmPFC lesions along with 46 healthy participants were assessed while taking the MST. Participants with vmPFC lesions exhibited excessive discrimination of lures by misidentifying them as foils (novel), indicating impaired mnemonic discrimination.

In a study by Alexander et al. [41], the performance of individuals with frontal lesions was compared to that of healthy participants in the CVLT. Researchers recorded the scores of each participant on every task of the CVLT while simultaneously recording their brain activity using FMRI. The results from the yes/no recognition task revealed that patients with frontal lesions exhibited significantly lower scores in the recognition test. The FMRI analysis revealed a high activity in the DLPFC among these patients. The lower scores were attributed to false-positive recognition of foils, indicating a defective semantic encoding of words leading to abnormal response bias.

Baldo et al. [7] conducted CVLT to compare the peformance of patients with frontal lesions to that of healthy participants. Three analyses were conducted in the yes/no recognition task: (1) the ability to differentiate between semantically related distractors, (2) the ability to differentiate between List B words with List A words and (3) the ability to differentiate unrelated novel words. The results showed that participants with frontal lesions showed poorer performance in the first two analysis compared to healthy participants, whereas both groups showed similar performance in the third analysis. This suggests that people with a lesioned prefrontal cortex use less semantic clustering ability leading to an increased tendency to endorse semantically related words.

Wais et al. [80] conducted an FMRI study to assess mnemonic discrimination, focusing on Ventrolateral Prefrontal Cortex (VLPFC). Participants were tasked with recognizing images as *old/new*. These images were taken from MST so that the images consisted of targets, novels and lures. To study how disruptions in VLPFC can impact mnemonic discrimination performance, disruption in neural activity of VLPFC was simulated in the participants. The results from this study showed that perturbations in VLPFC diminished discrimination of highly similar lures from targets indicating that VLPFC is necessary during successful mnemonic discrimination.

Johnson et al. [29] conducted a mnemonic discrimination test on rodents by disrupting the activity in the prefrontal cortex. The findings from this test suggested that when prefrontal cortex activity was disrupted, the rodents performance impaired significantly. Moreover, the results indicated that the degree of impairment depended on the target-lure similairity. Specifically, the performance was lower when the lures shared a feature overlap of 50% to 90% with the targets. These outcomes suggest the crucial role of the prefrontal cortex in resolving interference among stimuli, which is required for successful mnemonic discrimination.

Stuss et al. [70] conducted a comparison of word list learning performance between participants with bilateral frontal lobe damage and healthy individuals. The study comprised of a free recall task followed by a yes/no recognition task. The results from this study revealed significantly lower scores, specifically in the proportion of hits to correct rejections, among people with frontal damage.

These studies collectively highlight the involvement of different sub regions within the prefrontal cortex in recognition memory. In the next section, we will delve into studies that explore the interactions between the prefrontal cortex and hippocampus in the context of recognition memory.

Previous Studies on the Interactions between Prefrontal Cortex and Hippocampus in Recognition Memory and Mnemonic Discrimination

To understand how the hippocampus and prefrontal cortex collaborate in the encoding and retrieval of memory processes, it is important to understand the anatomical connections between these regions. Preston et al. [57] suggest that there exist pathways from the Medial Prefrontal Cortex (mPFC) to the hippocampus through the surrounding parahippocampal areas that facilitate the interaction between these regions during encoding and retrieval.

The study by Navawongse et al. [53] further strengthens the above hypothesis by studying the retrieval mechanism in rodents within a context-guided object association task by disabling the mPFC of the rodents. The results revealed that rodents with mPFC inactivation required more time to make choices and had reduced performance. In addition, the FMRI recordings during this task indicated that inactivation in mPFC led to reduction in the activation of hippocampal neurons, resulting in a failure to retrieve the required contextual information. These findings indicate the critical role of the prefrontal cortex in retrieving context based memories suggesting two way flow of information between the mPFC and hippocampus.

Wais et al. [79] conducted FMRI study focusing on medial temporal lobe and VLPFC while twenty subjects participated in an object yes/no recognition task. The study session involved the participants judging if the presented images (1) would fit inside a ladies medium shoe box? and (2) could be carried across the room using only the right hand? After the study session, participants had to judge if images presented were *old* or *new* based on what they had seen in the study session. The test images included targets, lures and novel items, similar to the MST. The recordings from the FMRI study revealed that discriminating highly similar lures from targets showed increased activity in the medial temporal lobe and VLPFC regions indicating the intact medial temporal lobe-cortical requirement for successful mnemonic discrimination.

King et al. [34] employed FMRI analysis to investigate various brain regions involved in the *recollection* process. They record the activity of the hippocampus, mPFC and other brain regions like parahippocampal cortex and left angular gyrus that are generally attributed to this strategic retrieval processes. Three different tasks were conducted: (1) Remember-know judgements of objects, (2) Associative memory procedure of pairs of objects and (3) source memory judgements, all of which involve recollection of memories in the brain. During the recollection process, the study revealed high levels of activations not only in the regions mentioned above but also in the DLPFC. The findings suggest that the functional co-ordination among these brain regions leads to successful recollection of memories.

Eichenbaum et al. [18] conducted an item recognition task on rodents, comparing those with hippocampal damage to those with mPFC damage. The results indicated that rodents with hippocampal damage tend to incorrectly identify *old* objects as *new* indicating a tendency towards forgetting memories. On the other hand, rodents with prefrontal cortex damage were inclined to identify *new* objects as *old* indicating a source memory impairment. These findings highlight the complementary roles of the prefrontal cortex and hippocampus in recognition tasks.

Understanding the roles of each of these brain regions in recognition paradigms helps us in designing the computational processes without compromising the biological plausibility. In the next section, we will present different AI algorithms that we utilize to implement our computational model.

2.5 Computational Modelling of Prefrontal-Hippocampal Interactions During Recognition Memory

For our proposed computational model, we have three key components. (1) Encoding the CVLT words as real-valued vector representations (2) designing hippocampus as an autoassociative attractor network (3) designing prefrontal cortex as a reinforcement learning agent. In this section, we discuss each of these AI algorithms.

2.5.1 Modeling CVLT Words as Vector Representations

To enable the model to perform the CVLT yes/no recognition task, we need to input the CVLT words as vector representations such that their semantic information is preserved [68, 47, 56]. For example, if we consider two semantically related words like spinach and cabbage, their corresponding vector representations should be close in the vector space. This representation of words is known as word embeddings, and it represents a significant breakthrough in the field of Natural Language Processing.

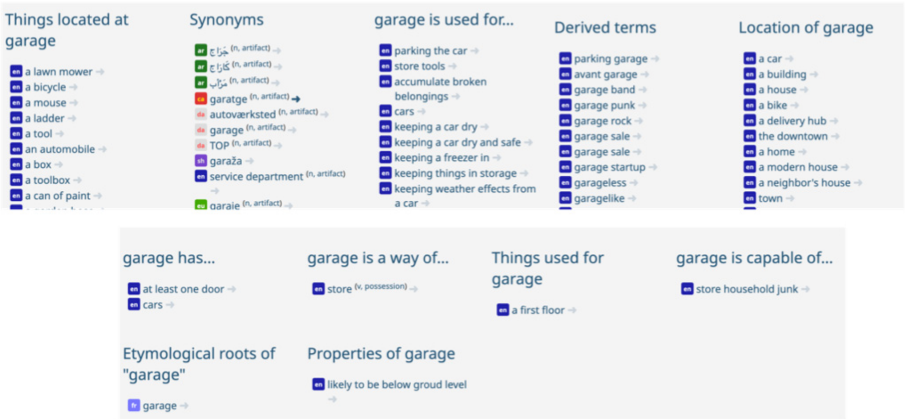
The main goal of training a neural network with word embeddings is to enable the network to comprehend words as closely as humans. This not only involves understanding words based on distances between word vectors (as shown in the above example of spinach and cabbage) but also encompasses various other dimensions [56]. As per the example in [56], the phrase *king is to queen as man is to woman* should be encoded in vector space by the equation *king* - *queen* $\propto$ *man* - *woman*. This encoding technique brings about clustering of words like *king* with *man* and *queen* with *woman*, creating precise analogical reasoning using vector representations [47, 56]. This type of vector representation captures the intrinsic details in a language aligning it more closely with human understanding.

Significant advancements in Natural Language Processing have introduced several techniques for word embeddings like distributed representations proposed by Mikolov et al. [47], GloVe by Pennington et al. [56] and ConceptNet by Speer et al. [68]. Each of these methods offer unique advantages compared to their predecessor. In our thesis we implement word embeddings using the ConceptNet embeddings proposed by Speer et al. [68]. Below, we explain in detail the ConceptNet embeddings and outline their advantages compared to other methods.

ConceptNet Embeddings

ConceptNet is a large multilingual knowledge graph that connects words and phrases of natural language [2]. The knowledge graph connects words and phrases called *terms* with labeled edges called *assertions*. For example, consider the phrase *a dog has a tail*. This phrase can be represented as a graph with start node: *dog*, end node: *tail* and a labeled edge: *HasA* [68].

ConceptNet is trained on diverse sources such as (1) Facts acquired from Open Mind Common Sense (OMCS) [67] and Games with a purpose [78] for common knowledge of language like phrases to express relationship between words, (2) Data from Wiktionary for multilingual vocabulary, (3) Open Multilingual WordNet [11] and JMDict [12] a Japanese multilingual dictionary, (4) OpenCyc [19] for commonsense knowledge, and (5) DBPedia [5] for facts extracted from Wikipedia sources. The combination of these sources creates a large multilingual knowledge graph containing 21 million edges and 8 million nodes [68]. This graph representation of the ConceptNet embeddings capture not just statistical patterns in large text corpora, but also explicit human-curated relationships and facts which makes it unique from the other word embedding techniques. This vast training source and graph representation of knowledge makes it superior to the other word embedding techniques. Fig. 2.1 represents an example of the representation of the word *garage* taken from ConceptNet. As seen in the figure, the ConceptNet contains not only the symmetric relationships like synonyms, antonyms and etymology but also asymmetric relationships of words like *used for* and *capable of* [68].



Things located at garage

- en a lawn mower
- en a bicycle
- en a mouse
- en a ladder
- en a tool
- en an automobile
- en a box
- en a toolbox
- en a can of paint

Synonyms

- ar جَرَاج (n, artifact)
- ar كَارَاج (n, artifact)
- ar مَرْآب (n, artifact)
- ca garatge (n, artifact)
- da autoværksted (n, artifact)
- da garage (n, artifact)
- da TOP (n, artifact)
- sh garaža
- en service department (n, artifact)
- eu garaje (n, artifact)

garage is used for...

- en parking the car
- en store tools
- en accumulate broken belongings
- en cars
- en keeping a car dry
- en keeping a car dry and safe
- en keeping a freezer in
- en keeping things in storage
- en keeping weather effects from a car

Derived terms

- en parking garage
- en avant garage
- en garage band
- en garage punk
- en garage rock
- en garage sale
- en garage sale
- en garage startup
- en garageless
- en garagelike

Location of garage

- en a car
- en a building
- en a house
- en a bike
- en a delivery hub
- en the downtown
- en a home
- en a modern house
- en a neighbor's house
- en town

garage has...

- en at least one door
- en cars

garage is a way of...

- en store (v, possession)

Things used for garage

- en a first floor

garage is capable of...

- en store household junk

Etymological roots of "garage"

- fr garage

Properties of garage

- en likely to be below groud level

Figure 2.1: Information about an English word *garage* as represented by ConceptNet

To assess the performance of ConceptNet, Speer et al. [68] compared the performance of ConceptNet with other word embedding techniques like distributed representations proposed by Mikolov et al. [47] and GloVe by Pennington et al. [56]. They evaluated ConceptNet's performance across a range of tasks like ranking word relatedness, choosing sensible ending to stories, and solving proportional analogies. In all these tasks ConceptNet consistently outperformed the other techniques. One such experiment in their research [68] is to rank the relatedness of word pairs and compare these rankings to actual human judgements. The ConceptNet outperformed other techniques in this task demonstrating its proficiency in representing the depth and breadth of word relationships [68].

2.5.2 Modeling the Hippocampus as an Autoassociative Memory

The Hippocampus region of the brain is involved in (1) rapidly encoding patterns (2) retrieving the encoded patterns when a partial cue is presented to it [21]. This forms the central idea that the hippocampus can function as an autoassociative memory [45, 46]. Therefore, in this section we introduce (1) Autoassociative memory, (2) their instantiation in Hopfield networks, and (3) Applications of Hopfield networks in AI.

Auto Associative Memory - Intuition Behind Hopfield Networks

Autoassociative memory, also referred to as content addressable memory, is a type of memory which can store patterns and is capable of retrieving them when a partial input is presented to it. The Hopfield network [25] is an artificial neural network that works as an auto associative memory.

Consider a pattern that is stored in memory. An ideal auto associative memory can retrieve this pattern error-free when a sufficient partial cue is presented to it [25]. To be able to achieve this, Hopfield [25] proposes the idea of a physical system which can be formulated as follows. Let us consider a physical system that has information stored in it. Let the different patterns stored in the physical system be denoted by $X = (x_1, x_2, ...x_N)$ which are local stable points in the system. Each element x_i in X is a vector representation of a pattern. Given these stored patterns, let us assume we have a pattern cue $y = x_1 + \Delta$ which is a point in the system nearer to x_1 and represents a partial cue for x_1. Any physical system that can reach the stable state x_1 from y can be known as content addressable memory or auto associative memory [25].

Classical Hopfield Network

The classical Hopfield network developed by Hopfield [25] designed the auto associative network as a sum of outer products of the patterns. Let us assume each of the pattern in $X = (x_1, x_2, ..., x_N)$ is of length d and $x_i \in \{-1, 1\}$. The weight matrix stores the patterns which models the Hebbian plasticity [25]

$$W = \sum_{1}^{N} x_i x_i^\top \tag{2.11}$$

To retrieve a stored pattern from the partial pattern y, we use a rule to update the states of the network until convergence is reached. The update rule in the classical Hopfield network is as follows:

$$y^{t+1} = sgn(Wy^t - b) \tag{2.12}$$

where $sgn()$ indicates the sign function. The convergence for this update rule is reached when $y^{t+1} = y^t$. Each time the update rule is converged, the energy function of the network is minimized.

$$E = -0.5y^{\top}Wy + y^{\top}b \tag{2.13}$$

If we tie back this network to the physical system mentioned above, the following is the intuition. If the weight matrix W is viewed as an energy landscape (the physical system), all the stored vectors in X are the local minima of this landscape. Therefore, during the retrieval process, if the partial cue y is considered a point in the energy landscape, the retrieval process is just a movement downhill towards the nearest local minimum which is, to one of the stored patterns in X that closely matches y.

The storage capacity of the Hopfield network is crucial as it determines the performance of the network. When the retrieval is error free, the classical Hopfield network has a storage capacity of [58]

$$C \approx \frac{d}{2log(d)}. \tag{2.14}$$

Ramsauer et al. [58] conducted an experiment where they store three images using the classical Hopfield network and try to restore the image by presenting a partial image to the network. They conclude that the retrieval is not error free when the images stored are strongly correlated. Thus, in conclusion, the major limitation of the classical Hopfield network is its low storage capacity.

Modern Hopfield Network (Dense Associative Memories)

The Modern Hopfield network proposed by Krotov et al. [38] is a discrete Hopfield network similar to the classical Hopfield network. However, the energy function is a polynomial interaction function which thereby increases the capacity of this network:

$$E = -\sum_{i=1}^{N} F(x_i^{\top}y) \tag{2.15}$$

where $F(z) = z^a$ is a polynomial interaction function with the degree of polynomial being a. Let $y[l]$ be the l^{th} component of the partial pattern y that is being updated. The update of the l^{th} component is defined as the difference of energy of the current pattern y and the next state when $y[l]$ is flipped. This component $y[l]$ is updated such that the energy is minimized. The update rule for the modern Hopfield network is

$$y^{new}[l] = sgn[-E(y^{(l+)}) + E(y^{(l-)})] \tag{2.16}$$

where $y^{(l+)}[l] = 1$ and $y^{(l-)}[l] = -1$. Unlike the classical Hopfield network, the modern Hopfield network does not have a weight matrix and so the energy function is the dot product of all stored patterns X with the state pattern y. The storage capacity of the modern Hopfield network is

$$C \approx \frac{1}{(2)(2a-3)!!} \frac{d^{a-1}}{log(d)} \tag{2.17}$$

In our proposed model, since we use ConceptNet to model the input which is a continuous valued representation of patterns, we need an autoassociative memory that supports continuous input. Ramsauer et al. [58] proposed a version of the modern Hopfield network for the continuous valued patterns.

Continuous Modern Hopfield Networks

Let us assume that the N stored patterns $X = (x_1, x_2, ..., x_N)$ are continuous patterns. The energy for the continuous Hopfield network is as follows

$$E = -lse(\beta X^\top y) + 0.5 y^\top y + \beta^\top log N + 0.5 \max(||M||_2) \tag{2.18}$$

where

$$M = \max_i ||x_i||$$

is the largest norm of all stored patterns and β is the inverse temperature. The update rule corresponding to the energy function is

$$y^{new} = X softmax(\beta X^\top y) \tag{2.19}$$

This update rule guarantees convergence of the energy function to a local minima.

The β value controls the learning dynamics of the network. If the patterns stored in the Hopfield network are different from each other, the convergence will be to a fixed point in the network which means it converges to one of the nearest stored patterns. If the patterns stored are similar to each other, the network converges to a metastable state which is close to the arithmetic mean of the stored patterns [58]. In such a case, high values of β, which corresponds to a low temperature can help each pattern to reach a stable state such that the attraction basin of each pattern remains separated from others. If the value of the β is lower, metastable states are formed due to which the retrieved pattern will be a combination of similar patterns in the network [58].

The important properties of this continuous Hopfield network according to Ramsauer et al. [58] are (1) global convergence to local minima (2) exponential storage capacity of $C \approx 2^{\frac{d}{2}}$, and (3) convergence after one update step.

Since the continuous Hopfield networks (1) handle continuous data, (2) are differentiable, and (3) converge in one step, they can be integrated into deep learning architectures. Ramsauer et al. [58] propose applications in deep learning where the Hopfield network can be integrated. In the next section we mention one such application of Hopfield network in deep learning.

Modeling Deep Networks Using Hopfield Networks - Application

The Update Rule of the Continuous Hopfield Network is Self Attention Mechanism in Transformers. In the study by Widrich et al. [83] it is shown that the update rule for the continuous Hopfield network is the self-attention mechanism of transformers [77]. Let us consider the update rule for the continuous Hopfield networks.

$$y^{new} = X softmax(\beta X^\top y) \tag{2.20}$$

Let us assume that we are updating multiple state patterns $Y = (y_1, y_2, ..., y_M)$ simultaneously instead of a single pattern. Therefore,

$$Y^{new} = X softmax(\beta X^\top Y) \tag{2.21}$$

We assume that the stored patterns X are keys (K) and state patterns Y are queries (Q). These are mapped into the Hopfield space with dimension d_k. Therefore we set

$$\hat{X}^\top = K = XW_K \tag{2.22}$$

$$\hat{Y}^\top = Q = YW_Q \tag{2.23}$$

$$V = XW_KW_V = \hat{X}^\top W_V \tag{2.24}$$

where Eq. (2.22), Eq. (2.23) and Eq. (2.24) are the Keys, Queries and Values respectively.

The dimensions of the matrices are $W_K \in \mathbb{R}^{d_x \times d_k}$, $W_Q \in \mathbb{R}^{d_y \times d_k}$, $W_V \in \mathbb{R}^{d_k \times d_v}$, $K \in \mathbb{R}^{N \times d_k}$, $Q \in \mathbb{R}^{M \times d_k}$ and $V \in \mathbb{R}^{N \times d_v}$

If $\beta = \frac{1}{\sqrt{d^k}}$, then for the update rule of the continuous Hopfield network, the self-attention is:

$$softmax(\frac{QK^T}{d_k})V = softmax(\beta YW_QW_K^\top X^\top)XW_KW_V \tag{2.25}$$

Having an attention mechanism with high storage capacity, this Hopfield layer can be used as a self-attention layer in deep learning architectures. Widrich et al. [83] propose a Deep Repertoire Classification (DeepRC) network with a self attention Hopfield layer which is tasked to predict immune status based on the immune repertoire sequences. Since each of these sequences are very large, finding patterns within those sequences can be a very challenging task. The study has shown that the DeepRC network outperforms other state-of-the-art methods despite the massive sequences.

2.5.3 Modeling Prefrontal Cortex as a Reinforcement Learning Agent

The prefrontal cortex region of the brain is said to be involved in higher cognitive tasks like planning, rule learning and reasoning [61]. It is hypothesized that the prefrontal cortex uses reinforcement learning to be able to perform such tasks. In this section we discuss (1) modeling higher cognitive tasks, and (2) reinforcement Learning

Modeling Higher Cognitive Tasks - Reward Prediction Error

Other than storing and retrieving memories, the brain is specialized in tasks like planning, grouping experiences into categories, and reasoning [65]. These are often referred to as higher cognitive tasks and are attributed to many brain areas like prefrontal cortex, basal ganglia and visual cortex [65]. In order to achieve these tasks, the brain is hypothesized to implement reward-driven learning. It is said that the brain structures like prefrontal cortex receive dopamine signals that convey reward prediction error which controls learning and facilitates goal directed behavior [54, 81, 73]. This idea of modulating actions through reward signals forms the basis for understanding how the prefrontal cortex works as a reinforcement learning agent.

Reinforcement Learning

Reinforcement Learning (RL) is a machine learning technique where an agent learns optimal behavior in an environment such that reward is maximized. The basic components in any RL problem are: (1) Environment - The world in which the agent operates, (2) State - Current situation of the environment, (3) Reward - Feedback signal from the environment (4) Policy - The strategy that an agent uses to maximize the reward (5) Value - Future rewards that can be expected based on the current action. Fig. 2.2 represents the basic action-reward loop of a RL problem

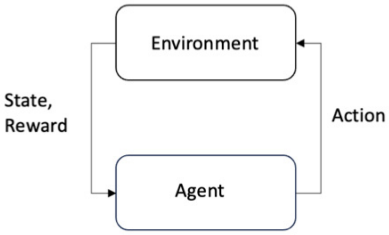


Figure 2.2: Action-Reward loop in Reinforcement Learning Problems.

In any RL problem, the agent interacts with the environment to achieve a goal. The agent does not have prior knowledge of what actions to take and so based on the reward signal and the state of the environment, the agent executes its actions so as to maximize the reward.

Let us assume that the agent starts interacting with the environment at a time step t. At this time step, the agent receives the state of the environment denoted by S_t. Based on the state of the environment, the agent selects an action A_t. The agent executes this action and one time step later (denoted by $t + 1$), it receives a reward signal from the environment R_{t+1} and the environment moves to the next state S_{t+1}. This loop continues until the agent reaches a terminal state, if one exists. The loop consists of a finite number of time steps and the entire process gives rise to a sequence of states, actions and rewards (S, A, R). The mathematical formalization of this RL loop is called a Markov Decision Process, and it forms the basis of formulating RL problems. Please note that, in this section we use the mathematical notation consistent with the formalization in Sutton and Barto [73].

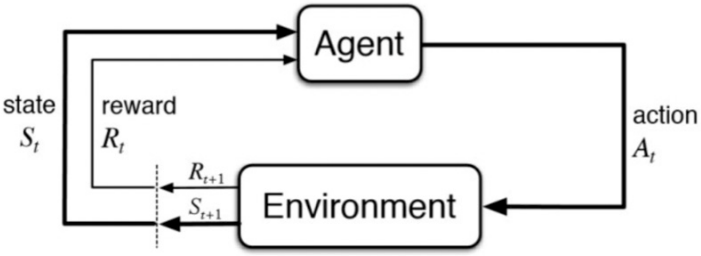


Figure 2.3: Interaction between the agent and environment in Markov Decision Process. (Source [73])

The variables R_t and S_t have probability distributions that are dependent only on the preceding state and action. The probability of $s' \in S$ and $r \in R$ occurring at time step t is dependent on the previous state $s \in S$ and action $a \in A$ occurring at the previous time step $t-1$.

$$p(s', r|s, a) = Pr\{S_t = s', R_t = r|S_{t-1} = s, A_{t-1} = a\} \tag{2.26}$$

Since the goal of the agent is to maximize the rewards, this is defined as *Discounted Expected Reward* denoted by

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3}.... \tag{2.27}$$

$$G_t = \Sigma_{k=t+1}^{T} \gamma^{k-t-1} R_k \tag{2.28}$$

The γ value is a discount factor which makes sure the rewards from immediate successive states weigh more than the rewards in the distant future.

Exploration vs Exploitation Tradeoff Since the agent has only limited knowledge of the environment, it should choose between taking actions that produced positive rewards in the past or exploring the environment to learn new information that may result in actions that can give better rewards in the future.

Value Function and Policy Every RL problem involves estimating value functions that approximate how valuable it is to take a certain action (also known as expected return) given the state of the environment. The value function is defined based on the policy the agent follows. The policy is typically the actions taken by the agent in its environment. Mathematically, a policy π is a mapping from states to probabilities of selecting actions.

$\pi(a|s)$ is the policy π the agent follows at any given time step t which is the probability of taking an action $A_t = a$ given the state $S_t = s$. Formally, the value function of a state $S_t = s$ under the policy π is defined as

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma_k R_{t+k+1}|S_t = s] \; \forall s \in S. \tag{2.29}$$

The $\mathbb{E}_\pi$ is the expected value given the agent follows the policy π. This is the state-value function. If we need to define the expected value of taking an action $A_t = a$ in a state $S_t = s$ under a policy π, it is called an action-value function denoted by $q_\pi(s, a)$

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma_k R_{t+k+1}|S_t = s, A_t = a] \tag{2.30}$$

Another property of the value function is the recursive relationships between the value of a state s to its successor states s'. Below is the Fig. 2.4 which represents this relationship. The head node is the current state s and the agent can take any action from state s (denoted by coloured circles) based on the policy π. Based on this the agent can land in any one of the successor states s' with probability p. The reward returned by the environment is denoted by r.

Mathematically, this can be defined as

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s] \tag{2.31}$$

$$= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \tag{2.32}$$

$$= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']] \tag{2.33}$$

$$= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')] \forall s \in S \tag{2.34}$$

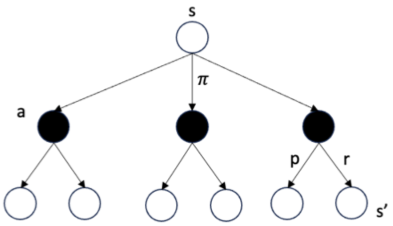


Figure 2.4: Relationship between values of current state with its successor states. (Source [73])

This equation is the sum of the probabilities of all future state, action, reward pairs weighted against the discounted value of the next state. This equation is called the Bellman equation.

Optimal Value Function and Policy To solve an RL problem, the goal is to find the optimal policy that can give maximum rewards. The basic idea behind this is that a policy π is said to be better than policy π', if the expected return of π is greater than the return for π'.

$$\pi \geq \pi' \iff v_{\pi}(s) \geq v_{\pi'}(s) \tag{2.35}$$

So every optimal policy will have a corresponding optimal value function.

$$v_*(s) = \max_{\pi} v_{\pi}(s) \tag{2.36}$$

is the optimal state-value function and

$$q_*(s,a) = \max_{\pi} q_{\pi}(s,a) \tag{2.37}$$

is the optimal action-value. The Bellman optimality equation for optimal state-value function can be written as:

$$v_*(s) = \max_{a} \sum_{s',r} p(s',r|s,a)[r + \gamma v_*(s')] \tag{2.38}$$

And for the action-value function can be written as

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)[r + \gamma q_*(s',a')] \tag{2.39}$$

Below is the Fig. 2.5 representing the optimal state-value and action-value selection

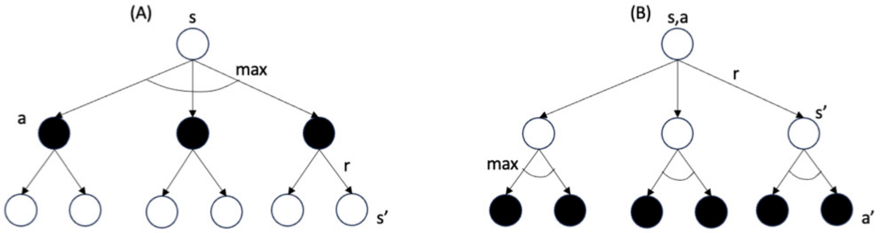


Figure 2.5: Pictorial representation for choosing **Panel (A):** $v*$ and **Panel (B):** $q*$ (Source [73])

Temporal Difference Learning Temporal Difference Learning algorithms are a type of RL algorithms that are used to model decision making processes in the brain [54]. FMRI images of the human brain have revealed similarities between phasic dopaminergic firing patterns and the characteristics of temporal difference reward prediction error [54]. In this section we will briefly discuss the Temporal Difference Learning and Q-learning algorithm that we use in our thesis to model the prefrontal cortex.

One of the problems in any environment is that the rewards are not immediately observable. So, in the TD methods, instead of calculating the total future reward, it estimates value based on immediate reward and the reward in the next time step without waiting for the final outcome. This method is called bootstrapping. Thus, TD learning is advantageous over other algorithms when the dynamics of the environment are unknown.

The value function for TD learning is defined as:

$$V(S_t) = V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)] \tag{2.40}$$

The $R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$ is the Temporal Difference Error which measures the difference between value at state S_t and the value estimate of the next time step $R_{t+1} + \gamma V(S_{t+1})$.

Q-learning Q-learning is a type of off-policy TD learning method. This is termed as off-policy because the action-value function is independent of the policy. The policy is only used to determine the state-action pairs that are visited. The action-value function of Q-learning is as follows:

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)] \qquad (2.41)$$

In the next section, we will review the existing computational models that attempted to design the computational processes in the brain for memory tasks. We will discuss the workings of these models, their limitations and how we overcome these limitations in our current model.

2.6 Existing Computational Models of Verbal Learning Tasks

In the first part of this section, we review the Temporal Context Model (TCM) [26] which is essentially an associative model of recall and recognition. Next, we will review the Becker and Lim [10] model which studies the dynamics of the verbal free recall in the context of the CVLT task by implementing a reinforcement learning agent along with associative memory that captures the recall strategies developed by the prefrontal cortex in the brain.

2.6.1 Temporal Context Model

The basic principle behind the TCM is *associate chaining theory* that conceptualizes the encoding process in the memory. In this theory, words are encoded such that they form associations with neighboring words, with associations becoming weaker for distant words. This is based on the *principle of recency* and *principle of contiguity* [26] observed in humans. Principle of recency is a phenomenon in memory where people tend to recall most recently heard information accurately. Principle of contiguity, on the other hand, refers to the phenomena where recall of one word facilitates recall of another word that was encoded closely in time to the previous word.

To model these effects Howard et al. [26] utilize the process of contextual coding during encoding and retrieval. A context can be defined as a randomly fluctuating signal that represents the current state of an event. For example, context can be a vector which can represent time. It can be thought of as a gradually drifting vector through a multidimensional context space. Because the context is gradually drifting, two items encoded next to each other have high overlap in their context vectors.

In TCM, the items are encoded along with context in the memory. This allows for overlapping contextual features for items that are encoded closely in time.

Let us assume a vector space F consisting of L vectors where each vector f_i in F corresponds to the item to be encoded. Another vector space T with each vector t_i which corresponds to the state of the context vector at time i for the corresponding item f_i. To associate the items and context two matrices are defined. One matrix is M^{TF} which forms the Hebbian outer product of both the context T and the corresponding item F. This matrix represents the strength of connections between each element in context T with each element in F.

$$M^{TF} = \sum_{i=1}^{L} f_i t_i^\top \tag{2.42}$$

Another associative matrix M^{FT} is formed as the Hebbian outer product of the item F and the corresponding context T. This matrix represents the strength of connections between each element in F with each element in T.

$$M^{FT} = \sum_{i=1}^{L} t_i f_i^\top \tag{2.43}$$

During the encoding process, when an item f_i is presented, two sequential steps occur. One, the item retrieves its existing context t_i^{IN} (if the item is repeated previously during the encoding). To retrieve this context, the item f_i will be presented to the M^{FT} which derives the context t_i^{IN}:

$$t_i^{IN} = M_i^{FT} f_i. \tag{2.44}$$

In the second step, this existing context t_i^{IN} is combined with the context in the previous time step t_{i-1} to update the current context t_i.

The context t_i for the item f_i is calculated as

$$t_i = \rho t_{i-1} + \beta t_i^{IN}, \tag{2.45}$$

where β is a free parameter. $0 \leq \rho \leq 1$ such that $||t_i|| = 1$ and determines the drift in the vector. Since the new context t_i is derived from the previous context t_{i-1}, the context vector changes gradually. Additionally updating the context based on this logic preserves the principle of recency and contiguity.

To satisfy the constraint of $||t_i|| = 1$, t^{IN} is also set to 1. Subsequently the association matrices M^{TF} and M^{FT} are updated. Fig. 2.6 depicts the recursive relationship between the item and context in the Temporal Context Model.

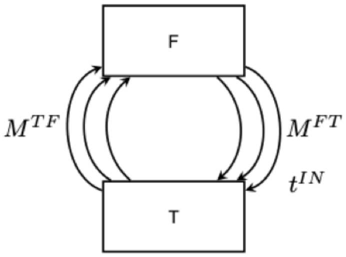


Figure 2.6: The recursive relationship between the Items and Context in TCM where one retrieves the other. Source([26])

During recall, the context serves as a cue to retrieve the item. Let the contextual cue be denoted by t. The item retrieved by the context t is

$$f^{IN} = M^{TF}t \tag{2.46}$$

The probability of recall of any item in the encoded vector space F is defined as based on the euclidean distance from f^{IN} to any vector f_i

$$P(f_i|f^{IN}) = \frac{\exp(-\frac{1}{\tau}||f^{IN} - f_i||^2)}{\sum_j \exp(-\frac{1}{\tau}||f^{IN} - f_i||^2)} \tag{2.47}$$

τ is a free parameter which controls the sensitivity of P_i to the differences in distances.

The matrix M^{TF} is cleared at the beginning of encoding each list. This approach of resetting the encoding list after completion of one study and recall session disables the model to recall items that are not part of the encoded list. The phenomena of recalling items that are not part of the encoding list is commonly observed in humans during free recall tasks. Therefore, the limitation of this model lies in clearing the M^{TF} after encoding and retrieval of each list.

Though the TCM focuses on providing computational intuition behind the storage and retrieval processes in the free recall paradigm, they do not explain the biological plausibility of these processes in the brain which is another limitation in this model.

In the next section we will review the Becker and Lim [10] model which focuses on developing a computational model that characterizes the biological aspect of the free recall paradigm in the brain.

2.6.2 Model of Prefrontal Cortex

The computational model developed by Becker and Lim [10] accounts for the interactions between the hippocampus and the prefrontal cortex in the free recall paradigm of verbal learning tasks. This is the only model which is built on the verbal free recall task of CVLT. This model particularly measures: (1) total words recalled across trials, (2) semantic clustering ability across trials, (3) effect of lesions in the prefrontal cortex in total words recalled and semantic clustering ability, and (4) advantage of blocked presentation of words in total words recalled and semantic clustering ability.

The model develops a strategic approach to learning words, enabling the self-organization of retrieval cues to perform the free recall task. The model consists of three layers. The first layer is the Lexical Semantic Memory which stores the CVLT words as binary representations. This layer is pretrained with a set of 100 English words including a mix of CVLT words, semantically related words that are not part of CVLT, and unrelated words. This feature allows the model to sometimes recall intrusions, overcoming the limitation in the TCM model and adding realism. The Medial Temporal Lobe is the second layer which is a binary autoassociator network. The third layer is the Prefrontal Cortex which acts as a reinforcement learning agent. The prefrontal cortex learns to recall words by developing mnemonic cues through trial and error. It receives a reward signal when correct words are recalled and a punishment signal if repetitions or intrusions are observed. This process allows the model to refine the retrieval strategies over the course of five trials of List A free recall.

Results from the model on the CVLT free recall task indicate that the lesions in prefrontal cortex reduces both the total correct responses and the semantic clustering scores. Additionally, the blocking of words during encoding enhances recall, even in the presence of prefrontal cortex lesions.

However, a notable limitation of the model is its binary representation of CVLT words. This binary representation of words will fail to capture the semantic information of English vocabulary as represented in the real world. For example, consider the words from *animals* category in the CVLT. The words *sheep* and *cow* should be closer in the vector space compared to each of their distances to the word *tiger*. Using binary encoding can make it challenging to capture these precise semantic distances between the words.

Since Becker and Lim [10] model is the only existing model in literature that studies the prefrontal-hippocampal interactions in CVLT task, we utilize this model to extend it to perform a yes/no recognition task. We also overcome the limitation of Becker and Lim [10] model by representing words as continuous word embeddings. Additionally, we adapt the intuition about the context representation and updation presented in the TCM model to design context in our proposed model. These extensions involve several modifications to the model and the next chapter delves into these details including the mathematical computations and experiments conducted in this study.

Chapter 3

Methodology and Experiments

This chapter explains the mathematical formalization of (A) the architecture of our proposed computational model, and (B) the experimental protocols used for model evaluation.

3.1 Model

Our model is an extension of a previous model by Becker and Lim [10], which is the only existing model in literature that depicts the interactions of prefrontal cortex and hippocampus on the free recall task of the CVLT, which is distinct from many free recall tasks in that it includes categories of related words. However, our work extends this model in several ways:

1. Whereas previous studies have modeled neural representations of words using categorically structured binary vectors, we represent the CVLT words using continuous word embeddings that are more likely to reflect the semantic distances between these words as perceived by humans [68].

2. Whereas previous studies have generally focused on binary autoassociative attractor networks such as classical Hopfield networks [25], we implement a medial temporal lobe hippocampal architecture capable of rapid storage and recall of continuous patterns using modern Hopfield networks [58].

3. We model temporal context by incorporating a context vector which consists of two features. (1) List context denoting the study list type and (2) Temporal context representing the time. The list context is a standardized binary vector which is constant for a study list and the temporal context is a slowly drifting vector generated by an Ornstein-Uhlenbeck Process. This allows us to better account for the fact that words on the CVLT are read in a defined sequence,

and allows our model to capture the element of temporal sequencing inherent in this task, whereas previous models did not account for this.

4. We modeled both the free recall components and yes/no recognition tasks in the CVLT to measure the overall recognition performance and the mnemonic discrimination performance. Therefore, rather than studying a model performing a single task as has previously been done (free recall), we are better simulating the fact that humans must perform multiple tasks during memory testing.

In this section, we will outline the structure of our computational model, which consists of three modules. Each module represents different parts of the brain which interact during free recall and recognition:

1. Lexical representation module
2. Medial temporal lobe module (MTL)
3. Prefrontal cortical module (PFC)

Fig. 3.1 represents the modules in the computational model

3.1.1 Lexicon

The lexical representation module consists of a function which accepts a token (a word), and returns a continuous embedding, which here models the activity of a set of d_w neurons. These embeddings represent high level semantic information about words, such that the embeddings of two words that mean similar things (e.g. *ox* and *bull*) will be closer in distance than the embeddings of two words that mean very different things (e.g. *ox* and *pencil*). To implement the word embeddings, we have used the state-of-the-art ConceptNet embeddings technique [68] which can capture semantic relationships between words as perceived by humans as opposed to abstract binary representations employed in previous studies. Fig. 3.2 illustrates a similarity matrix between ConceptNet embeddings of words found on the CVLT. Embedding similarity was measured using the dot product which is scaled to 0-1. If the dot product is higher (closely approaching to 1), the words exhibit strong semantic relation. We have replaced the word *subway* with *train* since, interestingly,

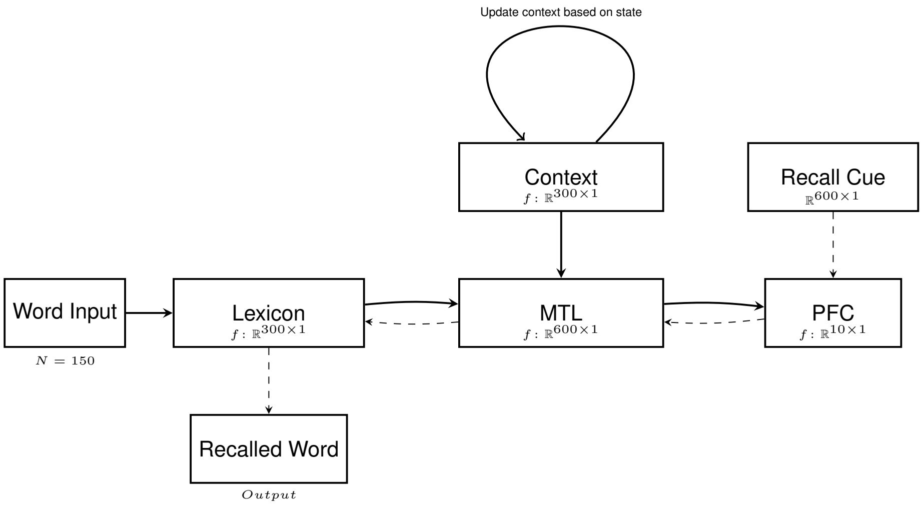

Update context based on state
Context
$f: \mathbb{R}^{300\times1}$
Recall Cue
$\mathbb{R}^{600\times1}$
Word Input
$N = 150$
Lexicon
$f: \mathbb{R}^{300\times1}$
MTL
$f: \mathbb{R}^{600\times1}$
PFC
$f: \mathbb{R}^{10\times1}$
Recalled Word
$Output$

Figure 3.1: Illustration of each layer in the model. Solid lines represent the encoding process and the dotted line represent the recall process. During the encoding process, each word is represented as a continuous word embedding in the lexicon. The MTL layer combines the embedding of the word from the lexicon with the current context. The self loop in the context module indicates that the context updates at each time step. MTL input is passed to the PFC, where it undergoes reinforcement learning. During the recall process, the recall cue is first passed to PFC which in turn passes the pattern cue to the MTL. The Hopfield network in the MTL converges the pattern cue to the desired pattern. This pattern is sent to the lexicon to generate the recall word.

the word subway in ConceptNet was not closely related to any words that fall under the *vehicles* category in CVLT.

Modeling the Lexicon and Transformation of a Word Token to Neural Representation

The Lexicon consisting of $N_w = 150$ words are modeled as a matrix,

$$A = \{a_{ij} : i = 1, 2, ..., d_w; j = 1, 2, ..., N_w\} \tag{3.1}$$

where the j^{th} column, $A_{:,j}$ is normalized to mean 0 and unit variance. The normalization of the word vectors did not result in any loss of information with respect to the semantic representation of the words. Fig. 3.3 depicts the PCA of CVLT word

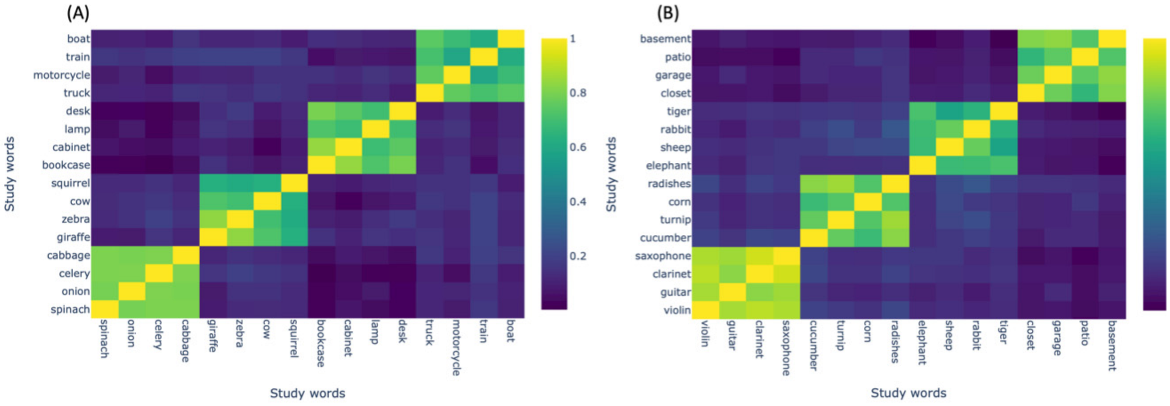


Figure 3.2: Heatmaps representing the dot product of CVLT words. The color bar represents the color range for the dot product which is scaled to 0-1. The words that are related the most have the highest dot product. **Panel (A):** Dot product between the CVLT List A words. **Panel (B):** Dot product between the CVLT List B words. The words in the graph are ordered based on their category and do not follow the original order in the CVLT test.

embeddings before and after normalization.

While the ConceptNet embedding set consists of 162298 English words, we restricted our lexicon to 150 words to facilitate computational efficiency. Furthermore, recalled CVLT words are generally common and often related to existing words on the study lists. As such, we selected the following four sets of words to comprise agents' lexicons:

1. **Set 1:** 48 words present in List A, List B and yes/no recognition test of CVLT.

2. **Set 2:** 30 new words which are from the same categories of CVLT List A and List B. Since the List A and List B words of CVLT are distributed among 6 different categories, we selected 5 new words from each category.

3. **Set 3:** 8 words are taken from two new categories fruits and devices which are not part of CVLT word categories. We choose 4 words from each of these categories.

4. **Set 4:** The remaining 64 words are random and are not related to any words selected above.

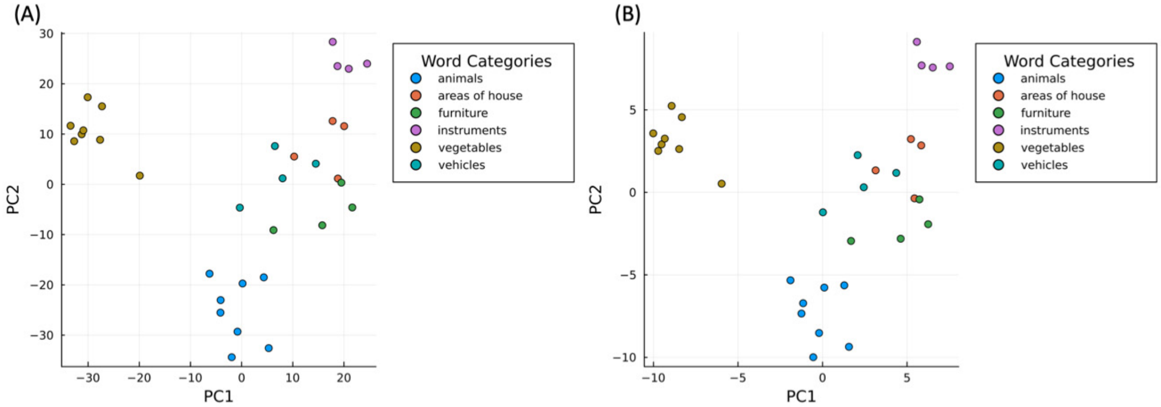


Figure 3.3: Principal Component Analysis (PCA) of the word embeddings. **Panel (A):** PCA of original ConceptNet embeddings. **Panel (B):** PCA of embeddings when Z-score normalization is applied. The Z-score normalization did not result in any loss of information in the word vectors

To generate the Set 2, Set 3 and Set 4 words for the lexicon, we used the LexOPS package in R [74]. The LexOPS package provides the capability to generate psycholinguistically controlled word stimuli. An inbuilt database of English words is also provided which contains a range of (1) lexical variables like parts of speech, word frequency (2) orthographic variables like bigram probability, (3) phonological variables like number of syllables, and (4) behavioral variables like proportion of people who know any given word, that are commonly used in psycholinguistic research.

Using the LexOPS library, we set various properties for words like (1) Zipf frequency, (2) Parts of speech (All CVLT words are nouns), (3) bigram probability, (4) Length of the words, (5) Frequency per Million words, (6) number of syllables, (7) level of concreteness (highly concrete words), and (8) familiarity (at least 98% know the word), to ensure the selected words are commonly used in the English language. For each of these conditions, we first recorded the values for the existing CVLT words (Set 1) using the LexOPS user interface and used the same range of values to select the other words. This is to ensure that all the words have the same properties of the CVLT words. Additionally, we conducted similarity checks by computing dot product between each selected word with the existing CVLT words to ensure the Set 2 words have high similarity, while Set 3 and Set 4 have low similarities with the Set 1 words.

Pre-Training the Lexicon layer

The Lexicon layer is first pre-trained via weights $D \in \mathbb{R}^{d_w \times N_w}$. The lexicon weights are initialized randomly and before the encoding process commences, the weights are trained through the entire 150 words in Lexicon which updates them to the corresponding word features in Lexicon [10].

$$D = D + \alpha_D A \tag{3.2}$$

where $0 < \alpha_D < 1$ is the learning rate.

Encoding in the Lexicon layer

During the encoding process, the weights in the Lexicon layer are trained via Hebbian learning which strengthens connections of the words encoded in the Lexicon.

$$D = D + \alpha_D (A \odot H) \tag{3.3}$$

where $H \in \mathbb{B}^{d_w \times N_w}$ such that the values are 1 for the embeddings of words that are encoded, and 0 otherwise. $0 < \alpha_D < 1$ is the learning rate which is the same as the pre-training learning rate.

Recall in the Lexicon layer

Recall in the Lexicon layer involves computing the dot product between the pattern generated by the hopfield network $\zeta_{k(1:d_w)}$ and D. The detailed description of choosing the recalled word through lexicon is described in Section 3.1.4.

$$i^* = argmax(\sigma(D^\top \zeta_{k(1:d_w)}))_i \tag{3.4}$$

3.1.2 Context

Since (A) words are presented to the agent sequentially over time, and (B) an essential aspect of the CVLT involves recalling words encountered in a particular context, it is imperative that our model includes neural representations of temporal and list-related contexts. The context representation consists of two features concatenated into a single vector. The first is a vector representation of the specific list with which a word was associated (i.e., List A), and the second is a vector representing time [4]. Let $C = (c_1, c_2, ...c_{N_{stored(t)}})^\top$ be the matrix of context representation $C \in \mathbb{R}^{d_c \times N_{stored(t)}}$ where $d_c = 300$ is the dimensionality of the context representation, and $N_{stored(t)}$ is the number of words that have been encoded by time t. The context vector c for each word is a concatenation of list type context vector L with dimensions $d_L = 250$ and temporal context vector T with dimensions $d_T = 50$

$$c = concat(L, T) \tag{3.5}$$

List Context

The first 250 elements in the context vector c denoted by L represent the type of list presented. It is created by first generating a binary vector where the first 125 elements are 1's for List A with the remainder set to 0. To ensure the concatenation of context with word embeddings have 0 mean and unit variance, the List A context vector is also standardized and scaled.

Temporal Context

The last set of 50 elements in the context vector c, denoted by T represents time as a slowly drifting vector with constant mean (0) and variance (1) generated by an Ornstein-Uhlenbeck Process [32]. This ensures that the items presented in close temporal succession have more similar contexts than items presented further apart [4]. The Ornstein–Uhlenbeck process is defined in continuous time which is a stochastic differential equation:

$$T_{t+1} = T_t + \theta(\mu - T_t)dt + \sigma dW_t \tag{3.6}$$

where $T_t \in \mathbb{R}^{d_T \times 1}$ is the drifting context vector for each time step t. The initial value for T_t is random and follows a standard normal distribution. dW_t is a Wiener process. The parameters μ, θ and σ are positive constants representing mean, mean reversion rate and volatility, respectively. This process has a stationary probability distribution due to the drift term $\theta(\mu - T_t)dt$. The drift term ensures that the process drifts towards the mean based on the mean reversion rate θ which makes it a mean-reverting process. When T_t is above mean, the process moves downward and when

T_t is below mean, it moves upwards. Fig. 3.4 below shows two simulations of the Ornstein-Uhlenbeck Process.

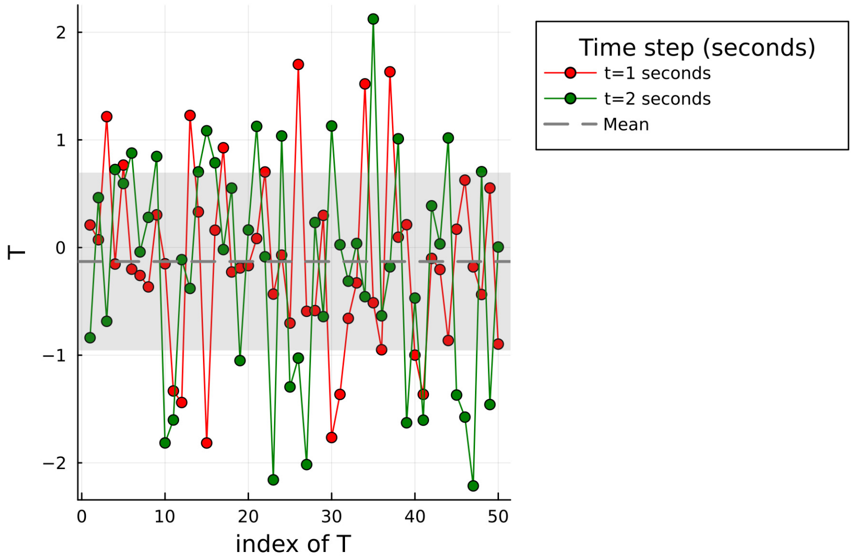


Figure 3.4: Two illustrative simulations of a one-dimensional Ornstein-Uhlenbeck Process with $\mu = 0$ and $\sigma = 1$

We set dt to 1 second, such that $T_t - T_{t+1}$ represents the change in context representation over 1 second. During the immediate free recall phase, each word is read out to participants at a rate of no more than 2 seconds per word. As such, for successive words presented to the agent, the context vector will be incremented by 2 which represents an increment of 2 seconds.

Upon completion of the encoding phase of each immediate free recall trial, temporal context is incremented by $dt = 2$ seconds, and the recall phase is commenced. After recall of each word, temporal context is again incremented by $dt = 2$ seconds.

Before the start of the Long Delay Free yes/no recognition process, there will be approximately a 20 minute delay. So, the temporal context is incremented by 1200 seconds to represent this delay. After each step in the long delay yes/no recognition process is incremented by 2 seconds.

3.1.3 Medial Temporal Lobe Memory System

The Medial Temporal Lobe (MTL) represents the hippocampus in the brain modeled based on the Modern Hopfield Network for continuous patterns [58] with d_{mtl} units where $d_{mtl} = d_w + d_c$.

Encoding in the Medial Temporal Lobe

Encoding the words into MTL constitutes storage of $N_{stored}(t)$ embeddings of words observed by the agent. For example, during Immediate Free Recall A, we present $N_{stored}(t) = 16$ words such that $Y = (y_1, y_2, ...y_{N_{stored}(t)})^\top$.

These are concatenated with context such that each encoded word and its context is represented as $s = (y, c)^\top$ where $s \in \mathbb{R}^{d_{mtl} \times 1}$.

$$M = \begin{pmatrix} Y \\ C \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{N_{stored}(t)} \end{pmatrix}^\top = \begin{pmatrix} y_1 & c_1 \\ y_2 & c_2 \\ \vdots & \vdots \\ y_{N_{stored}(t)} & c_{N_{stored}(t)} \end{pmatrix}^\top \tag{3.7}$$

For the first encoding trial, the embedding and context are concatenated as shown above. From the second encoding trial, the existing temporal context of each word in the M is averaged with the temporal context of the current trial. This models the typical observation in human recognition memory studies that recency and contiguity association of the words is higher in the first trial and gradually decreases in the following trials [32]. That is, humans typically recall words in the sequence in which they were heard on the first recall attempt. As the number of encoding trials and recall attempts increases, the serial order of words tends to become less prominent in the order in which subjects recall words. For example, consider the i^{th} encoded word s_i. During the first trial, the temporal context of the encoded word s_i is T_t where t is the current time step. When the second trial commences, the time step progresses to $t = t + n_{steps}$ where $n_{steps} = (N_{stored}(t) - i + N_{attempts}) \times 2$. (Note: We multiply by 2 because encoding or recall of each word takes 2 seconds each). Let the temporal context during this time step be $T_{current}$. The corresponding temporal context of s_i is updated to $T_t = (T_t + T_{current})/2$

Attention During Encoding

In the verbal learning tasks, participants often attend more to the initial words presented in the task, with attention decreasing towards the middle and end of the list [32]. As per Kahana [32], the context associated with each word can explain the recency effect but the high attention towards the early list of words is independent of the recency effect. This study also mentions that this phenomenon is the source of the primacy effect observed during the free recall. Sederberg et al. [64] modeled attention as an exponentially decaying scalar value. That is, for the i^{th} encoded word, the attention is modeled as follows:

$$\phi_i = \phi_s \exp(-\phi_d(i-1)) + 1 \tag{3.8}$$

where $\phi_s + 1$ denotes the value of attention for the first encoded word. ϕ_d denotes the decay in the attention for the rest of the encoded words. The Fig. 3.5 below represents the decay in the value of attention ϕ for each of the encoded words.

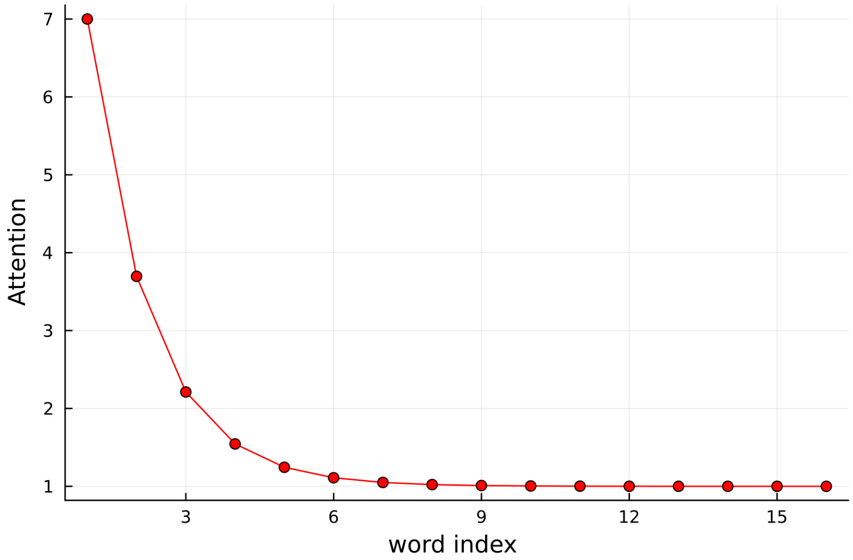


Figure 3.5: The exponential decay in the attention parameter ϕ for each of the encoded words in List A. The parameters are $\phi_s = 6.0$ and $\phi_d = 0.8$

During the first trial of encoding, we incorporate this attention parameter by multiplying the scalar value ϕ to each of the embeddings before adding the context.

Modeling attention ensured a learning curve in the total number of words recalled during the immediate free recall.

For i^{th} encoded word y_i, we multiply the corresponding attention value ϕ_i as shown below.

$$y_i = \phi_i \odot y_i \tag{3.9}$$

Recall in the Medial Temporal Lobe

The recall process in the modern Hopfield network of MTL is implemented as follows:

$$\zeta^{'} = Msoftmax(\beta M^{\top}\zeta) \tag{3.10}$$

where ζ is an embedding of the cue generated during the recall phase. $M \in \mathbb{R}^{d_{mtl} \times N_{stored}(t)}$ is the set of previously memorized words at any time t. $N_{stored}(t) =$ 16 List A words stored during Immediate Free Recall A (IFRA). β is the inverse temperature [58]. The detailed description of recall process in the MTL is described in Section 3.1.4

3.1.4 Inhibition of Return (Response Suppression)

Response suppression is a method where a temporary suppression mechanism is added for the recently recalled words so that it prevents the model from recalling the same words repeatedly [10, 32]. This phenomenon is known as inhibition of return [35]. In the Becker and Lim [10] model, the experiment lacking inhibition of return made many repetition errors and recalled fewer words and had shallow learning curves. To prevent this, we have included inhibition of return in the Lexicon and Medial Temporal Lobe layers.

Response suppression at Medial Temporal Lobe

To implement response suppression, a short term memory window $STM = (j_1^*, j_2^*, ..., j_{stm}^*)$ is maintained during the recall process. j^* is the index of the previously recalled word in the set of memorized words M. The capacity of the STM is $stm = 7$ which means any time during the recall process, the STM holds the indices of the last recalled words. Choosing the STM capacity $stm = 7$ helped in reducing the repetitions and improved the overall recall responses across all the agents

This STM is used to create a mask $Z \in \mathbb{B}^{d_{mtl} \times N_{stored(t)}}$ where

$$Z_{i,j} = \begin{cases} 0 & at \quad j \in STM \\ 1 & elsewhere \end{cases} \tag{3.11}$$

During the recall process in the medial temporal lobe, we include the mask Z in the Hopfield network equation. This selectively disables the previously recalled words from the encoded words M. We then allow the network dynamics of the Hopfield network to reach equilibrium by running

$$\zeta_1 = (Z \odot M) softmax(\beta (Z \odot M)^\top \zeta_0) \tag{3.12}$$

$$\zeta_2 = (Z \odot M) softmax(\beta (Z \odot M)^\top \zeta_1) \tag{3.13}$$

and so on, until convergence is reached at time k.

$$\zeta_k = (Z \odot M) softmax(\beta (Z \odot M)^\top \zeta_{k-1}) \tag{3.14}$$

such that the final converged pattern is ζ_k.

In our model, the convergence criteria for the Hopfield network is that the norm of the previous state pattern and the current state pattern is less than the tolerance level of 0.01.

$$\sqrt{\sum_{i=1}^{d_w+d_c} (\zeta_{k_i} - \zeta_{(k-1)_i})^2} < 0.01 \tag{3.15}$$

Response suppression at Lexicon

In cases where the model recalls intrusions, it is not possible to execute the response suppression of these intrusions in the medial temporal lobe. So, the inhibition of return mechanism is also included in the lexicon so that the this mechanism also applies to intrusions that are not encoded in the M matrix. The response suppression logic for lexicon is similar to medial temporal lobe.

The STM is used to create a mask $Z \in \mathbb{B}^{d_w \times N_w}$ where

$$Z_{i,j} = \begin{cases} 0 & at \quad j \in STM \\ 1 & elsewhere \end{cases} \tag{3.16}$$

During the recall process in the lexicon, we include the mask when selecting the recalled word as follows:

$$i^* = argmax(\sigma((D \odot Z)^\top \zeta_{k(1:d_w)}))_i \tag{3.17}$$

where $(D \odot Z)^\top \zeta_{k(1:d_w)}$ is the dot product between the retrieved Hopfield pattern and the embeddings stored in lexicon. We compute the dot product so that we obtain the word in lexicon that closely matches with the retrieved Hopfield pattern. We scale the dot product values using the softmax function $\sigma((D \odot Z)^\top \zeta_{k(1:d_w)})$ and select the index of the word i^* that has the highest dot product. The word corresponding to this index i^* in the lexicon is the word recalled.

Let the word recalled be denoted by $\hat{x}$. If the model is allowed to always select the word corresponding to the highest dot product, we observed that the model is recalling only from the encoded words. To ensure the model recalls out of list words (intrusions) in some cases, we incorporated a non-greedy policy of choosing the word unit. Rather than consistently choosing the word with the highest dot product, we sometimes let the model choose the word with the next highest dot product. To implement this, we first mask the word $\hat{x}$ in the Lexicon D.

$$D_{i,j} = \begin{cases} 0 & at \quad j \in i^* \\ D_{i,j} & elsewhere \end{cases} \tag{3.18}$$

Using the updated Lexicon weights D, we calculate the dot product of the Hopfield pattern with D and select the word unit with the highest dot product.

$$i^* = argmax(\sigma((D \odot Z)^\top \zeta^*_{1:d_w}))_i \tag{3.19}$$

3.1.5 Prefrontal Cortex

The Prefrontal Cortex (PFC) layer is modeled following Becker and Lim [10] as a RL agent using Q-learning that learns iteratively through trial and error, developing strategies to facilitate the recall process. In essence, this model learns abstract representations of word categories by reinforcement learning. The PFC layer is represented as $Q = (Q_1, Q_2, ...Q_{d_{pfc}})^\top$ where $Q \in \mathbb{R}^{d_{pfc} \times 1}$ and $d_{pfc} = 10$.

Encoding in the Prefrontal Cortex

During the encoding phase, the PFC agent takes in each of the encoded words from the medial temporal lobe $s = (y, c)^\top$ where $s \in \mathbb{R}^{d_{mtl} \times 1}$ and calculates state action values, called Q-values, for each unit of PFC. It is the linear sum of weighted inputs from the MTL layer and the bias b which stores the inputs from previous actions [10]. Q-values for each PFC unit at time step t is calculated as:

$$Q_t = F_t^\top s_t + b_t \tag{3.20}$$

where $F \in \mathbb{R}^{d_{mtl} \times d_{pfc}}$ is the matrix of connection weights from the MTL to the PFC. We have that

$$b_t = \delta_t^{slow} b_t^{slow} + \delta_t^{fast} b_t^{fast}, \tag{3.21}$$

where $b^{slow} \in \mathbb{R}^{d_{pfc} \times 1}$ and $b^{fast} \in \mathbb{R}^{d_{pfc} \times 1}$ are the learnable slow bias and fast bias terms in the PFC at time step t. As per Becker and Lim [10], the fast bias and slow bias are added to the prefrontal cortex layer to maintain the prefrontal units' activity over time during the encoding and recall trials. Maintaining the previous activity through the bias terms enables sustained response in the prefrontal units and allows the model to perform semantic clustering where consistent activation in one prefrontal unit during recall corresponds to recalling words of the same category together [10]. This strategy is inspired from the FMRI study by Constantinidis et al. [14] where sustained activations were observed in DLPFC neurons of primates during working memory tasks. At each time step, the b_{slow} parameter accumulates the history of actions taken at all previous time steps of one encoding and recall trial whereas b_{fast} only includes the action of the previous time step [10]. If the model lacks bias, it exhibits very low semantic clustering scores across all trials [10]. δ_t^{slow} and δ_t^{fast} are learnable scalars that determine the strength of the slow and fast bias.

The larger the Q-value, it is highly likely that the model will select that action such that the rewards are maximized. Based on Becker and Lim [10], instead of letting the model always choose the highest Q-value, we express the Q-values as probabilities using the softmax function. This allows for some randomization during the action selection. Thus, the Q-value with the highest probability is most likely chosen but it is not always guaranteed to be chosen (we denote this as $\max_{soft}$ in the equation below). This lets the model balance between exploration and exploitation.

The optimal action value (Q^*) is caculated as

$$Q^* = \max_{soft}(\sigma(Q_i)) = \max_{soft}(\frac{e^{Q_i}}{\sum_{i=1}^{d_{pfc}} e^{Q_i}}) \quad for\ i = 1, 2, \ldots, d_{pfc} \tag{3.22}$$

$$p = \arg\max_i(Q_i = Q^*) \tag{3.23}$$

is the prefrontal unit corresponding to the optimal Q-value selected.

Both the weights and bias are learnable parameters which are trained using Q-learning:

$$F_{t+1} \leftarrow F_t + \alpha_s(r_t + \gamma max(F_t^\top s_{t+1} + b_t) - P^\top(F_t^\top s_t + b_t)) \odot s_t P^\top \tag{3.24}$$

where $F_t \in \mathbb{R}^{d_{mtl} \times d_{pfc}}$ is the matrix of state-action values corresponding to weights from MTL to PFC at the current time step t. P is a one hot vector taking value of 1 at the index p using Eq. (3.23). In the encoding phase, the model only focuses on the current word to be encoded and does not consider any future actions. As a result we set the γ value to 0. The $r_t + \gamma max(F_t^\top s_{t+1} + b_t) - P^\top(F_t^\top s_t + b_t))$ is called the reward prediction error denoted by $error_t$ at a time step t. r_t is the reward at time t is the feedback received by the agent based on the action taken. Since the model does not generate any response during the encoding phase, the reward is set to 1 [10]. $0 < \alpha_s < 1$ is the learning rate.

The connection weights from PFC to MTL layers denoted by $B \in \mathbb{R}^{d_{pfc} \times d_{mtl}}$ is also trained using the same logic described above:

$$B_{t+1} \leftarrow B_t + \alpha_s(r_t + \gamma max(B_t s_{t+1} + b_t) - P^\top(B_t s_t + b_t)) \odot P s_t^\top \tag{3.25}$$

The bias terms are updated as follows.

$$b^{slow} = \lambda \cdot b^{slow} \tag{3.26}$$

$$(b_p^{slow})_{p \in P} = P \cdot b^{slow} + (1 - \lambda) Q^* \tag{3.27}$$

where p is the index of the optimal Q-value, Q^* from Eq. (3.23) and λ is a scalar value which defines the strength of past activity to be accumulated.

$$(b_p^{fast})_{p \in P} = Q^* \tag{3.28}$$

which is reset after each time step t and hence it is called fast bias. The δ_{slow} and δ_{fast} values are updated at each time step by Q-learning:

$$\delta_{t+1}^{slow} \leftarrow mean(diag(\delta_t^{slow} I + \alpha_s^{\delta}(r_t + \gamma max(F_t^{\top} s_{t+1} + b_t) - P^{\top}(F_t^{\top} s_t + b_t)) \odot b_t^{slow} P^{\top})) \tag{3.29}$$

and

$$\delta_{t+1}^{fast} \leftarrow mean(diag(\delta_t^{fast} I + \alpha_s^{\delta}(r_t + \gamma max(F_t^{\top} s_{t+1} + b_t) - P^{\top}(F_t^{\top} s_t + b_t)) \odot b_t^{fast} P^{\top})) \tag{3.30}$$

where $I \in \square^{d_{pfc} \times d_{pfc}}$ and $\gamma = 0$

Recall in the Prefrontal Cortex

When recall is initiated, the first retrieval cue to the PFC is $s_{Nstored(t)}$ which is the last word studied during encoding. This is based on the recency effect in free recall phenomenon where participants are more likely to recall the last studied word first [31].

From the next recall attempt, the recall cue will be the word recalled in the previous attempt. If the previous recalled word is part of the encoded word list (M), then the embedding along with its context stored during the encoding is presented as the next recall cue. If the recalled word is an intrusion, the context is set to zero.

This retrieval cue is used to select the PFC unit p that produces the optimal action value using

$$Q_t = F_t^{\top} s_{Nstored(t)} + b_t \tag{3.31}$$

In the first recall attempt, the F_t and b_t are the weights and biases from the last studied word in the encoding phase. The optimal action value (Q^*) is selected using the same softmax function using the same policy defined in the encoding

$$Q^* = \max_{soft}(\sigma(Q_i)) = \max_{soft}(\frac{e^{Q_i}}{\sum_{i=1}^{d_{pfc}} e^{Q_i}}) \quad for\ i = 1, 2, \ldots, d_{pfc} \tag{3.32}$$

$$p = \arg \max_i (Q_i = Q^*) \tag{3.33}$$

This is used to generate the MTL pattern cue as follows:

$$\zeta = (B_t^{\top} P) \odot Q^* \tag{3.34}$$

which is the weighted sum of activations from the PFC layer. $B \in \mathbb{R}^{d_{pfc} \times d_{mtl}}$ is the PFC to MTL weight matrix. P is a one hot vector taking value of 1 at the index p. Q^* is the optimal action value for the corresponding retrieval cue of PFC.

This MTL pattern ζ is standardized and scaled and then transmitted to MTL which uses equation Eq. (3.12) to converge the Hopfield network to an equilibrium and generate the final converged pattern ζ_k. The standardizing and scaling of the prefrontal output is important since the patterns stored in the Hopfield network and the embeddings stored in the Lexicon are all standardized and scaled. We calculate the dot product between this converged pattern and Lexicon weights D to generate the recalled word using Eq. (3.17). We pass this resultant pattern along with its context (s_{t+1}) as the retrieval cue for the next recall attempt.

The weights and biases are learnable parameters that are trained at each time step t during recall, in addition to their training during encoding:

$$F_{t+1} \leftarrow F_t + \alpha_r \eta (r_t + \gamma max(F_t^\top s_{t+1} + b_t) - P^\top (F_t^\top s_t + b_t)) \odot s_t P^\top \tag{3.35}$$

where $F_t \in \mathbb{R}^{d_{mtl} \times d_{pfc}}$ is the matrix of state-action values corresponding to weights from MTL to PFC at the time step t. P is a one hot vector taking value of 1 at the index p which is the index of the optimal Q-value, Q^* from Eq. (3.33). The $r_t + \gamma max(F_t^\top s_{t+1} + b_t) - P^\top (F_t^\top s_t + b_t)$ is called the reward prediction error denoted by $error_t$ at a time step t. r_t is the reward at time t, which represents the feedback received by the agent based on the previous word recalled. If the word recalled is the correct word, then the reward is set to 1. If the word is not from the list (intrusion) or if it is repeated, the reward is set to -1. γ is the discount factor. η is the learning rate set to 1. If there are repetitions during the recall, it is set to 0.25 [10].

$$B_{t+1} \leftarrow B_t + \alpha_r \eta (r_t + \gamma max(B_t s_{t+1} + b_t) - P^\top (B_t s_t + b_t)) \odot P s_t^\top \tag{3.36}$$

The bias terms b_{slow} and b_{fast} are updated as follows.

$$b^{slow} = \lambda \cdot b^{slow} \tag{3.37}$$

$$(b_p^{slow})_{p \in P} = P \cdot b^{slow} + (1 - \lambda) Q^* \tag{3.38}$$

$$(b_p^{fast})_{p \in P} = Q^* \tag{3.39}$$

The δ_{slow} and δ_{fast} values are updated at each time step by Q-learning algorithm

$$\delta_{t+1}^{slow} \leftarrow mean(diag(\delta_t^{slow} I + \alpha_r^{\delta} \eta (r_t + \gamma max(F_t^\top s_{t+1} + b_t) - P^\top (F_t^\top s_t + b_t)) \odot b_t^{slow} P^\top)) \tag{3.40}$$

and

$$\delta_{t+1}^{fast} \leftarrow mean(diag(\delta_t^{fast} I + \alpha_r^{\delta} \eta (r_t + \gamma max(F_t^\top s_{t+1} + b_t) - P^\top (F_t^\top s_t + b_t)) \odot b_t^{fast} P^\top)) \tag{3.41}$$

Since each recall trial in the CVLT test ends when the agent cannot remember any more words, our model ends a recall trial when two consecutive words are recalled incorrectly.

3.2 Implementation of the Agent Under a Yes/No Recognition Paradigm

Let $Y^* = (y_1^*, y_2^*, ..., y_N^*)$ be the embeddings of the test list for a recognition model. These are concatenated with List A context such that each encoded word is represented as $s^* = (y^*, c^*)^\top$. The context c^* is set to the List A context.

We propose Approach 1, detailed in Section 3.2.1, which involves making recognition memory judgements that involves full recollection process by engaging both the prefrontal cortex and medial temporal lobe. This full recollection process encompasses performing reinforcement learning through the prefrontal cortex as described in Section 3.1.5 and subsequently passing the output from the prefrontal cortex to the medial temporal lobe. This allows for the pattern completion process as described in Section 3.1.3. After the full recollection, our approach is designed to perform recognition memory judgements.

In addition to approach 1, we also implemented two other approaches in Section 3.2.2 and Section 3.2.3. The second approach only uses the executive control system of the reinforcement learning process as described in Section 3.1.5 to make recognition memory judgements and the third approach only utilizes the recollection through the medial temporal lobe to make recognition memory judgements. This detailed analysis allows us to evaluate the significance of full recollection in recognition process and unravel the individual contributions of the prefrontal cortex and medial temporal lobe.

3.2.1 Approach 1: Full Recollection

The first approach implements full recollection, where each recognition cue undergoes the entire recollection process similarly to what occurs during recall. This idea of recollection is discussed by Mandler in [43] and Yonelinas in [85] where the recognition judgement process is viewed as a strategic search process that retrieves context information of the cue. This idea is also proposed by Kahana in [32] where the recognition process is modeled by presenting a cue and retrieving its associated pattern, similar to cued recall. Fig. 3.6 below outlines our step by step process for this approach.

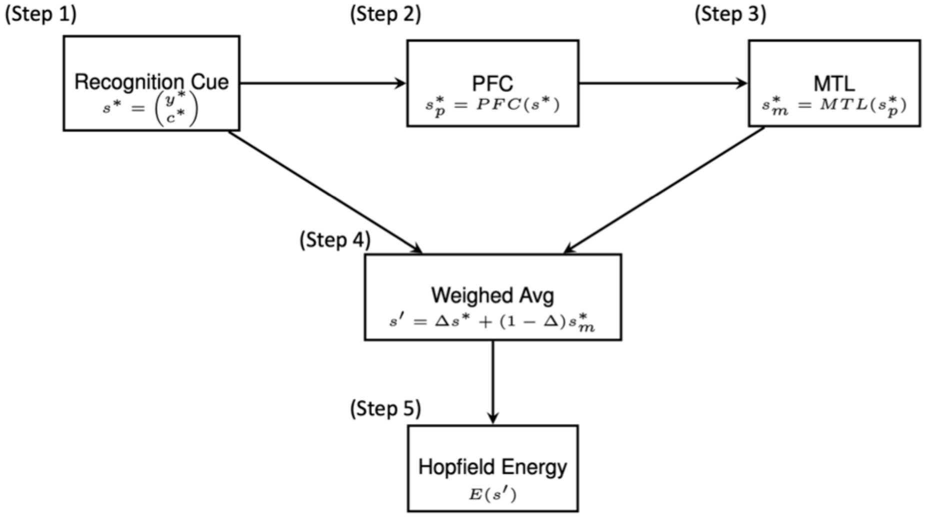


Figure 3.6: The step by step process of the first approach to the yes/no recognition model. **Step 1:** The initial recognition cue s^* is the embedding of the CVLT recognition word combined with its context. **Step 2:** The recognition cue is passed through the PFC layer to generate the PFC cue s_p^*. **Step 3:** The output from PFC layer s_p^* is passed through the MTL layer, which is allowed to perform iterative pattern completion, in order to generate s_m^* and complete the full recollection process. **Step 4:** The weighted average of the original embedding and the retrieved pattern is then calculated to generate s'. **Step 5:** Hopfield energy of the resultant pattern using Eq. (2.18) is calculated.

In step 1 we add context to the recognition word embedding to create $s^* = (y^*, c^*)^\top$. In step 2, we present s^* as a cue to the prefrontal cortex module which is denoted by $s_p^* = PFC(s^*)$ in Fig. 3.6. The $PFC(s^*)$ involves passing the recognition cue through the PFC to generate the MTL pattern cue using the Eqs. (3.31) to (3.34). The step 3 denoted by $s_m^* = MTL(s_p^*)$ converges the MTL pattern cue in the Hopfield network to generate the final converged pattern as shown in Eq. (3.12). These three steps constitute the recollection process and are common for both recognition and recall. In Step 4, the converged pattern is averaged with the original recognition cue to generate s' using the weighted average method. The weight variable Δ is a scalar value which is between 0 and 1. We measured the performance of the recognition test, by varying the values of Δ.

In Step 5 we calculate the energy of the Hopfield network using Eq. (2.18) as presented in [58] for each of these patterns. As per the formalization of energy function in [58], each pattern stored in the Hopfield network represents a fixed point in the network leading to minimization of the energy value. So, when the retrieved pattern in our model closely resembles the patterns stored in Hopfield network M, the energy associated with the pattern will be minimum. When the energy value of a pattern is minimum, it is highly likely that the pattern is stored in the memory M. Therefore, we use this energy value to estimate the probability of recognizing the pattern as *old* or *new*.

3.2.2 Approach 2: Recollection through PFC

In the second approach each recognition cue does not undergo full recollection; rather the output from the PFC layer is combined with the original embedding for recognition. As per Becker and Lim [10], in contrast to the MTL which stores the detailed episodic trace of the patterns (embedding with context), the PFC layer stores the representations as a mnemonic code enhancing particular attributes like categorical properties of the patterns to facilitate systematic retrieval and enable semantic clustering. Based on this idea, we propose that the PFC layer typically enhances the categorical properties of the cue which when combined with the recognition cue, helps in recognizing a pattern as *old* or *new*. Comparing this approach with the previous one enables us to test the importance of full recollection as opposed to relying on this partial recollection process in recognition judgments. The Fig. 3.7 below outlines our step by step process for this approach.

In step 1, we add context to the recognition word embedding to create $s^* = (y^*, c^*)^\top$. In step 2, we present s^* as a cue to PFC which is denoted by $s_p^* = PFC(s^*)$

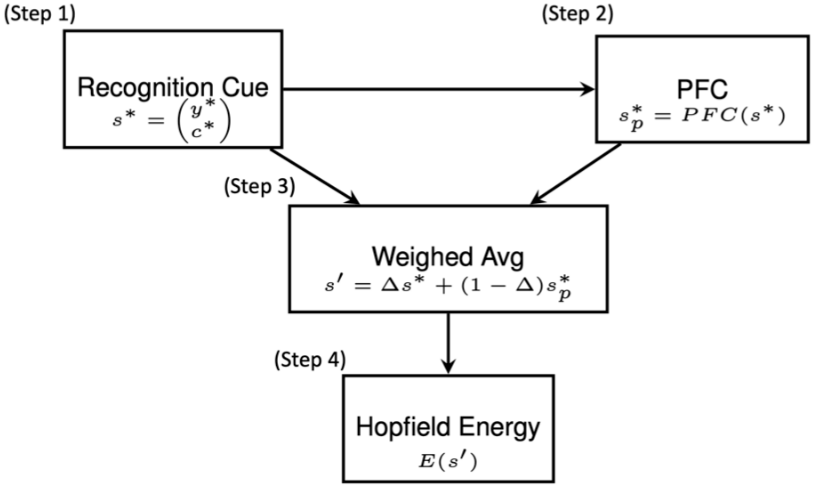


Figure 3.7: The step by step process of the second approach to the yes/no recognition model. **Step 1:** The initial recognition cue s^* is the embedding of the CVLT recognition word combined with its context. **Step 2:** The recognition cue is passed through the PFC layer to generate the PFC cue s_p^*. **Step 3:** The weighted average of the original embedding and the pattern from PFC layer is calculated to generate s'. **Step 4:** Hopfield energy of the resultant pattern using Eq. (2.18) is calculated.

in the Fig. 3.7. The $PFC(s^*)$ involves passing the recognition cue through the PFC to generate the MTL pattern cue using the Eqs. (3.31) to (3.34) present in the Recall section of the Prefrontal Cortex. In step 3, the pattern generated from PFC layer is averaged with the original recognition cue to generate s' using the weighted average method. The weight variable is a scalar value Δ which is between 0 and 1. We measured the performance of the recognition test, by varying the values of Δ. In step 4 we calculate the energy of the Hopfield network as in Approach 1. This energy is used as a measure for probability of recognizing the pattern as *old* or *new*.

3.2.3 Approach 3: Recollection through MTL

In the third approach the recognition process does not involve the PFC layer. The recognition cue is directed to the MTL layer to perform pattern completion and then is combined with the original embedding for recognition. Since the MTL is always attributed to recognition process and mnemonic discrimination, comparing this approach with approach 1 allows us to compare how pattern completion process in the hopfield network contributes to recognition as opposed to full recollection. The Fig. 3.8 below outlines our step by step process for this approach.

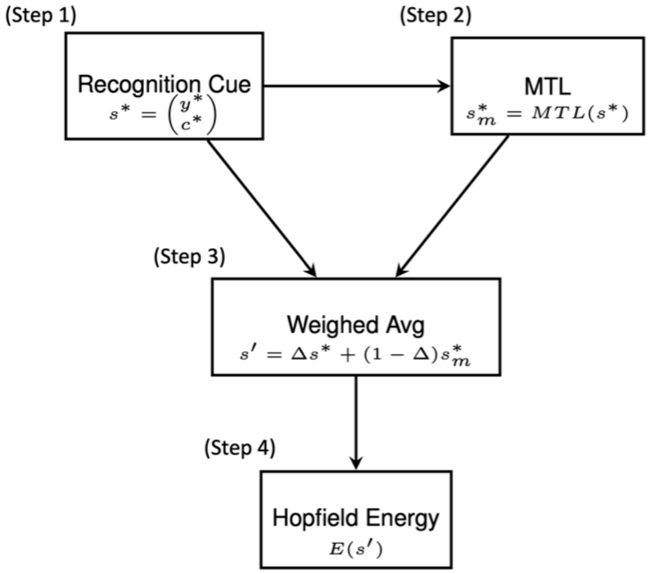


Figure 3.8: The step by step process of the third approach to the yes/no recognition model. **Step 1:** The initial recognition cue s^* is the embedding of the CVLT recognition word combined with its context. **Step 2:** The recognition cue is passed through the MTL layer to generate the MTL cue s^*_m. **Step 3:** The weighted average of the original embedding and the pattern from MTL layer is calculated to generate s'. **Step 4:** Hopfield energy of the resultant pattern using Eq. (2.18) is calculated

In step 1, we add context to the recognition word embedding to create $s^* = (y^*, c^*)^\top$. In step 2, we present s^* to the Hopfield network which converges the word embedding to generate the final converged pattern from MTL as shown in Eq. (3.12). In step 3, the converged pattern is averaged with the original recognition cue to generate s' using the weighted average method. The weight variable is a scalar value Δ which is between 0 and 1. We measured the performance of the recognition test, by varying the values of Δ. In step 4 we calculate the energy of the Hopfield network same as presented in approach 1. This energy is used as a measure for probability of recognizing the pattern as *old* or *new*.

3.3 Experiments

We carried out series of experiments to assess performance of the free recall and recognition tasks. Initially we aimed to validate the free recall results obtained by extending our baseline model by Becker and Lim [10]. The details of this extended model can be found in Section 3.1. Our evaluation of the free recall performance included assessing the metrics of the CVLT: total number of correctly recalled words and semantic clustering scores as outlined in Section 2.3.1.

In the second phase of our experiments, we introduced the delayed yes/no recognition task. We evaluated the results of delayed mnemonic discrimination and overall recognition performance on this task using a new measure of mnemonic discrimination proposed by Leger et al. [40] which is described in Section 2.3.2.

To provide a basis for comparision, we measured the performance of our extended free recall model against the Becker and Lim [10] with a focus on the first three variables: group, blocked and trial, presented in Table 3.1. Furthermore, we evaluated the yes/no recognition performance across the variables: group, blocked and delta, presented in Table 3.1.

Sl No.	Variable	Values	Description
1	group	intact/lesion	A flag indicating if the agent is intact or lesioned
2	blocked	blocked/unblocked	A flag indicating if the CVLT words are presented in the order of their categories during encoding or not
3	trial	between 1-5	Indicating one of the 5 trials executed during free recall
4	delta (Δ)	between 0-1	The weight variable that is utilized to calculate the weighted average during yes/no recognition task

Table 3.1: Different scenarios for measuring model performance. The first two variables group and blocked are used in testing both free recall and recognition performance. The variable trial only applies to free recall since the free recall process takes place over a series of 5 trials. The delta (Δ) value is only applicable for yes/no recognition test where it is used to calculate the weighted average.

This section has a detail explanation of (1) different tasks simulated, (2) different agents for the experiment, and (3) statistical analysis for the free recall and mnemonic discrimination experiments.

3.3.1 Verification of Becker and Lim Results on Free Recall

Specific Task Instantiation

Our first experiment was to extend our baseline model by Becker and Lim [10] as mentioned in the Section 3.1. We calculated CVLT metrics (1) total number of correct recalled words and (2) semantic clustering scores and compared them to the results reported in the Becker and Lim model [10].

As in Becker and Lim [10], our experiments involved simulating two kinds of agents, intact and lesioned. As our study focuses on prefrontal cortex lesions, we disabled 30% of nodes in the PFC layer and along with its incoming and outgoing connections to MTL. Disabling more than 30% of the connections did not result in any reduction in the number of correctly recalled responses. We then compared the correct responses and semantic clustering scores of the lesioned model with the intact model, where none of the connections were disrupted.

In the next task, we extended the previous experiment by evaluating how a blocked presentation of CVLT words affects model performance. To do this, we simulated free recall experiments by encoding the words in the order of their categories. This experiment was carried out for both intact and lesioned agents.

Simulated Agents

Our experiments were conducted on four different agent types, intact, lesion, intact blocked and lesion blocked. Within each of these groups, we executed the free recall experiment on 100 agents. We then calculated the total number of correct recalled words and semantic clustering scores (refer Section 2.3.1) for each trial and averaged the results within each group. This allowed us to get an overall performance measure of each group and make comparisons among the different groups.

Statistical Analysis

Total Number of Correct Recalled Words To estimate the effect of the first three variables mentioned in Table 3.1 on the number of correct responses $\hat{C}$ generated during free recall, we employed the linear regression model described below (presented in R syntax):

$$\textbf{Model 1: } \hat{C} \sim group + blocked + trial + group : blocked \tag{3.42}$$

Observed Semantic Clustering Scores To assess the influence of the above mentioned variables on the observed semantic clustering scores OS generated during free recall, we conducted linear regression model described below:

$$\textbf{Model 2: } OS \sim group + blocked + trial + group : blocked \tag{3.43}$$

Both the models examined the impact of each of these variables and also analyzed the interaction between the group and blocked variables on the number of correct responses generated and semantic clustering scores respectively.

3.3.2 Evaluation of Model Performance on Delayed Mnemonic Discrimination

Specific Task Instantiation

In our second experiment, we introduce the yes/no recognition task following the completion of the free recall task. For this recognition task, we calcuated the metrics MDI and REC (Section 2.3.2) for each of the three approaches detailed in Sections 3.2.1 to 3.2.3. Similar to the free recall task, we applied these three approaches to different agent types: intact, lesion, intact blocked and lesion blocked. We then conduct a comparitive analysis of the MDI and REC for each of these agent types.

Simulated Agents

Our experiments were conducted on four different groups: intact, lesion, intact-blocked and lesion-blocked. Each group was subjected to the experiment with 100 agents. For each individual agent, we consider a range of Δ values to identify the influence Δ value on the MDI and REC. The Δ is the weight variable utilized in computing the weighted average during the recognition process (refer Figs. 3.6 to 3.8).

In our assessment we fit a five-parameter logistic function for each agent as described in the Eq. (2.7), to evaluate the model's capability to classify new words as a function of distance. Additionally, we calculated the MDI (Eq. (2.8)) and REC index

(Eq. (2.9)) to assess how the mnemonic discrimination and recognition performance vary when semantically related words are present, as described in Leger et al. [40].

Logistic Mnemonic Discrimination Index (MDI)

We fit the sigmoidal function in Eq. (2.7) to each agent, categorized by their group (intact/lesion), blocked (blocked/unblocked) and delta values. We utilized the non-linear least squares in the LsqFit.jl package of the Julia programming language (Dhanyaasri et al., (manuscript in preparation)). After determining the best fitted curve for each of the groups, we calculated the MDI using Eq. (2.8).

Statistical Analysis To estimate the effects of the variables group, blocked and delta on MDI, we performed linear regression analysis described below:

$$\begin{aligned} \textbf{Model 4: } MDI \sim group + blocked + delta + group : blocked + \\ group : delta + blocked : delta + group : blocked : delta \end{aligned} \tag{3.44}$$

Recognition Index (REC)

After calculating the MDI, we calculated the REC using Eq. (2.9) for each of the agents and compared the recognition performance across the group, blocked and delta variables.

Statistical Analysis To estimate the effects of the variables group, blocked and delta on REC, we performed linear regression analysis described below:

$$\textbf{Model 6: } REC \sim group + blocked + delta + group : blocked \tag{3.45}$$

$$\begin{aligned} \textbf{Model 6: } REC \sim group + blocked + delta + group : blocked + \\ group : delta + blocked : delta + group : blocked : delta \end{aligned} \tag{3.46}$$

For all the linear regression models mentioned above, we conducted the statistical power calculations to assess whether the sample sizes in these experiments were adequate to draw meaningful and statistically significant conclusions. All power calculations, Tables C.1, C.2, D.1 to D.3 and E.1 to E.6 show a power of 1 for significance level of 0.001 showing that all the models yield results with high reliability for our sample size of 100 agents. The next chapter will present the results for all the experiments described in this section.

Chapter 4

Results

4.1 Verification of Becker and Lim Results on Free Recall

Fig. 4.1 shows the number of correct responses and observed semantic clustering scores for different group of agents and blocked presentation of input over the course of five trials. As reported in the Becker and Lim [10] results, lesioning nearly one third of connections in the prefrontal cortex layer and its interactions with the hippocampus degraded the recall performance. The effect of lesioning on total number of correct recalled words shows a significant negative impact ($\beta = -1.75, p < 0.001$; Table 4.1). The prefrontal cortex lesions also has a significant negative impact on semantic clustering scores ($\beta = -1.45, p < 0.001$; Table 4.2).

Further, as reported in Becker and Lim [10], we see a significant negative effect of unblocked list presentation on semantic clustering scores ($\beta = -0.32, p = 0.006$; Table 4.2). Although the unblocked presentation of words shows a negative impact for semantic clustering scores, we did not observe a significant impact on the total number of correct recalled words. Furthermore, no significant impact was observed for the interaction between unblocked and lesion as reported by Becker and Lim [10].

4.2 Comparison of Model Performance on Delayed Mnemonic Discrimination

4.2.1 Logistic Mnemonic Discrimination Index (MDI)

Approach 1: Full Recollection

Based on Fig. 4.2 below, the MDI value is highest for lower delta values (i.e. with more PFC involvement) indicating that MDI heavily relies on the recollection process involving both prefrontal cortex and medial temporal lobe. Linear Regression

	Dependent variable (Total Recalled Words)		
Predictors	**Estimates**	**CI**	**p**
(Intercept)	11.2	10.96 – 11.44	<**0.001**
blocked [unblocked]	-0.22	-0.45 – 0.01	0.064
group [lesion]	-1.75	-1.98 – -1.52	<**0.001**
trial	0.2	0.14 – 0.26	<**0.001**
blocked [unblocked] × group [lesion]	-0.07	-0.40 – 0.26	0.660
Observations	2000		
R2 / R2 adjusted	0.202 / 0.201		

Table 4.1: Results of the linear regression model Eq. (3.42)

	Dependent variable (Semantic Clustering)		
Predictors	**Estimates**	**CI**	**p**
(Intercept)	4.3	4.06 – 4.54	<**0.001**
blocked [unblocked]	-0.32	-0.56 – -0.09	**0.006**
group [lesion]	-1.45	-1.68 – -1.22	<**0.001**
trial	0.31	0.26 – 0.37	<**0.001**
blocked [unblocked] × group [lesion]	0.08	-0.24 – 0.41	0.616
Observations	2000		
R2 / R2 adjusted	0.169 / 0.167		

Table 4.2: Results of the linear regression model Eq. (3.43)

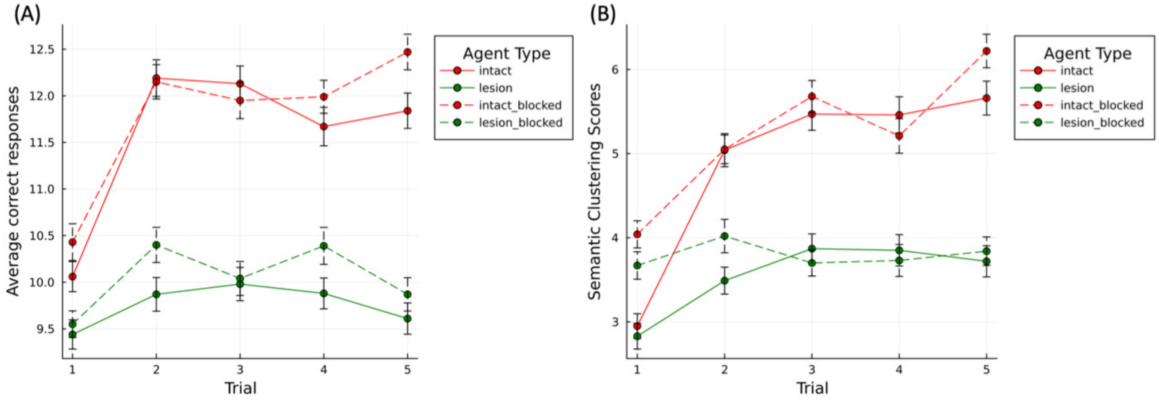


Figure 4.1: List A Immediate Free Recall results. **Panel (A):** Total number of correct recalled words over the course of five trials averaged over each group. **Panel (B):** Semantic clustering scores over five trials averaged over each group

results show significant negative effect of delta on MDI ($\beta = -0.42, p < 0.001$; Table 4.3). Furthermore, lesioning the model has significant negative effects on the MDI ($\beta = -0.07, p = 0.004$; Table 4.3). These results collectively indicate that MDI performance is associated with full recollection and benefits from the intact prefrontal cortex.

Approach 2: Recollection through PFC

Based on Fig. 4.3, the effect of delta on MDI is consistent with the results seen in Approach 1 (Section 4.2.1). Linear Regression results in Table 4.4 show a significant negative impact of delta on MDI ($\beta = -0.33, p < 0.001$).

Approach 3: Recollection through MTL

Based on Fig. 4.4, delta is the only significant variable with linear regression results showing a statistical significance of ($\beta = -0.63, p =< 0.001$) shown in Table 4.5. This emphasizes on the importance of hippocampal recollection in mnemonic discrimination.

	Dependent variable (MDI)		
Predictors	**Estimates**	**CI**	**p**
(Intercept)	0.83	0.79 – 0.86	<**0.001**
blocked [unblocked]	0	-0.05 – 0.05	0.928
group [lesion]	-0.07	-0.12 – -0.02	**0.004**
delta	-0.42	-0.48 – -0.37	<**0.001**
blocked [unblocked] × group [lesion]	0.07	0.00 – 0.14	**0.036**
blocked [unblocked] × delta	0.01	-0.07 – 0.09	0.838
group [lesion] × delta	0.09	0.01 – 0.17	**0.025**
(blocked [unblocked] × group [lesion]) × delta	-0.11	-0.22 – 0.00	0.057
Observations	4153		
R2 / R2 adjusted	0.158 / 0.156		

Table 4.3: Linear regression model results Eq. (3.44) for Approach 1

	Dependent variable (MDI)		
Predictors	**Estimates**	**CI**	**p**
(Intercept)	0.65	0.62 – 0.67	<**0.001**
blocked [unblocked]	0.02	-0.01 – 0.05	0.28
group [lesion]	0.03	-0.00 – 0.06	0.06
delta	-0.33	-0.37 – -0.29	<**0.001**
blocked [unblocked] × group [lesion]	0.03	-0.02 – 0.07	0.262
blocked [unblocked] × delta	-0.02	-0.07 – 0.04	0.579
group [lesion] × delta	-0.03	-0.09 – 0.02	0.255
(blocked [unblocked] × group [lesion]) × delta	-0.05	-0.12 – 0.03	0.228
Observations	4307		
R2 / R2 adjusted	0.245 / 0.244		

Table 4.4: Linear regression model results Eq. (3.44) for Approach 2

	Dependent variable (MDI)		
Predictors	**Estimates**	**CI**	**p**
(Intercept)	0.9	0.88 – 0.91	<**0.001**
blocked [unblocked]	0	-0.02 – 0.02	0.773
group [lesion]	-0.01	-0.02 – 0.01	0.574
delta	-0.63	-0.65 – -0.60	<**0.001**
blocked [unblocked] × group [lesion]	0.01	-0.02 – 0.04	0.529
blocked [unblocked] × delta	0	-0.03 – 0.03	0.9
group [lesion] × delta	0	-0.03 – 0.03	0.867
(blocked [unblocked] × group [lesion]) × delta	-0.01	-0.05 – 0.04	0.753
Observations	4397		
R2 / R2 adjusted	0.730 / 0.729		

Table 4.5: Linear regression model results Eq. (3.44) for Approach 3

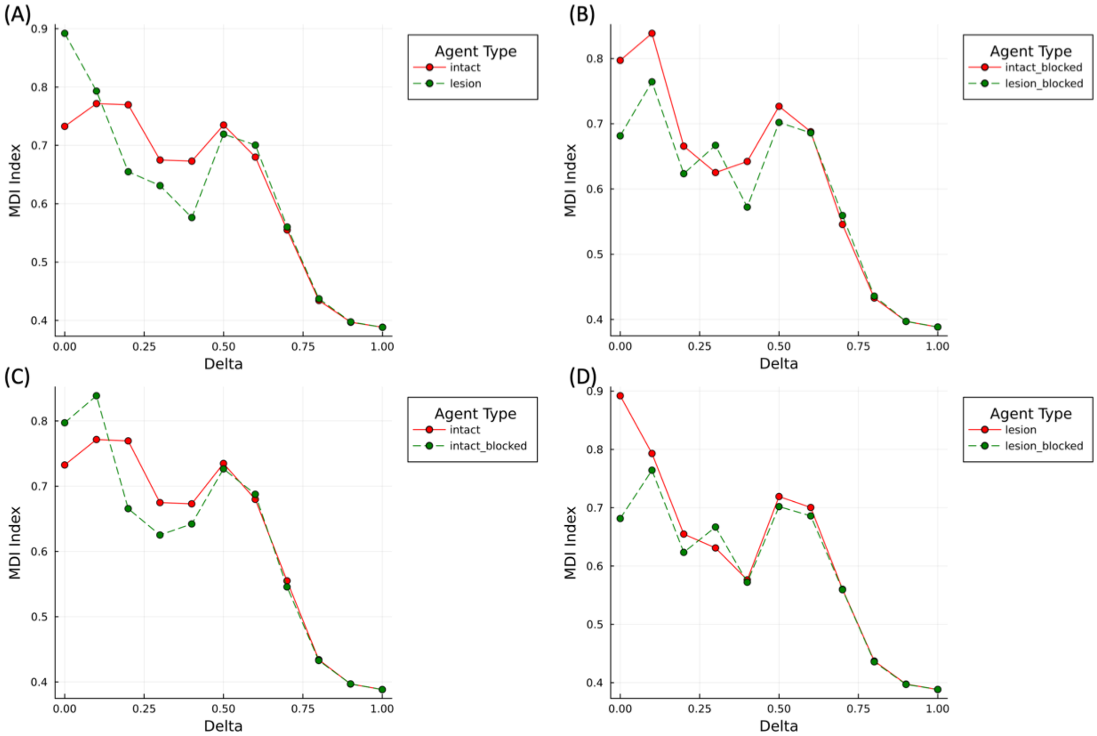


Figure 4.2: The MDI vs delta for **Approach 1**. **Panel (A):** The MDI vs delta for intact agents vs lesion agents. **Panel (B):** The MDI vs delta for intact agents vs lesion agents when the input is a blocked presentation. **Panel (C):** The MDI vs delta for intact agents when the words are encoded in blocked vs unblocked fashion. **Panel (D):** The MDI vs delta for lesion agents when the words are encoded in blocked vs unblocked fashion

4.2.2 Recognition Index (REC)

Approach 1: Full Recollection

Fig. 4.5 represents the REC vs Delta plots. Notably, the delta value has a significant positive effect of REC, indicating that REC is compromised when recollection takes place. This is in contrast to MDI where recollection enhances the mnemonic discrimination performance. Linear Regression results of delta on REC are ($\beta = 0.75, p < 0.001$; Table 4.6) and ($\beta = 0.77, p < 0.001$; Table 4.7)

Furthermore, when delta interactions are not introduced, the two way interaction model Eq. (3.45), shows that lesioning the prefrontal cortex has a statistically significant negative effect on the REC ($\beta = -0.02, p = 0.014$; Table 4.6).

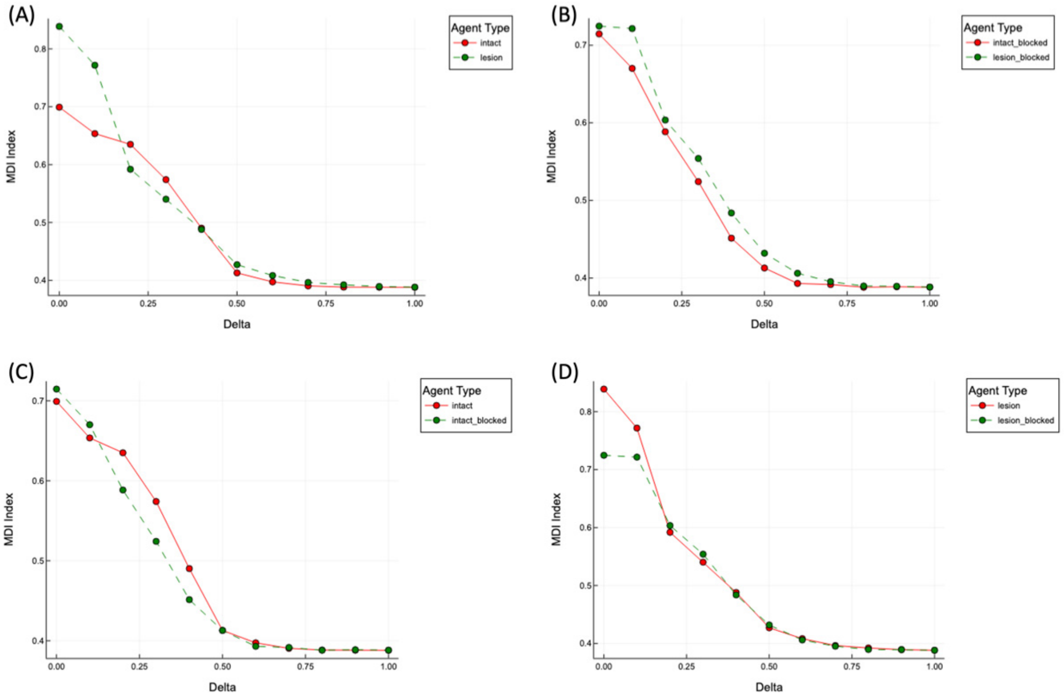


Figure 4.3: The MDI vs delta for **Approach 2**. **Panel (A):** The MDI vs delta for intact agents vs lesion agents. **Panel (B):** The MDI vs delta for intact agents vs lesion agents when the input is a blocked presentation. **Panel (C):** The MDI vs delta for intact agents when the words are encoded in blocked vs unblocked fashion. **Panel (D):** The MDI vs delta for lesion agents when the words are encoded in blocked vs unblocked fashion

Unblocked list presentation has a significant positive impact on the REC ($\beta = 0.02, p = 0.005$; Table 4.6) and ($\beta = 0.07, p < 0.001$; Table 4.7) unlike MDI which did not see any effect of blocking.

Approach 2: Recollection through PFC

Based on Fig. 4.6 below, the effect of delta value is consistent with the results seen in approach 1 (Section 4.2.2). Linear Regression results indicate a significant positive impact of the delta value ($\beta = 0.63, p < 0.001$; Table 4.8), ($\beta = 0.59, p < 0.001$; Table 4.9) on the recognition performance.

Prefrontal cortex lesions further impact REC showing a statistically significant

	Dependent variable (*REC*)		
Predictors	**Estimates**	**CI**	**p**
(Intercept)	0.12	0.10 – 0.13	<**0.001**
blocked [unblocked]	0.02	0.01 – 0.04	**0.005**
group [lesion]	-0.02	-0.04 – -0.00	**0.014**
delta	0.75	0.73 – 0.77	<**0.001**
blocked [unblocked] × group [lesion]	-0.03	-0.05 – -0.00	**0.019**
Observations	4153		
R2 / R2 adjusted	0.619 / 0.619		

Table 4.6: Linear regression model results Eq. (3.45) for Approach 1

	Dependent variable (*REC*)		
Predictors	**Estimates**	**CI**	**p**
(Intercept)	0.11	0.08 – 0.13	<**0.001**
blocked [unblocked]	0.07	0.04 – 0.10	<**0.001**
group [lesion]	-0.02	-0.05 – 0.01	0.162
delta	0.77	0.73 – 0.81	<**0.001**
blocked [unblocked] × group [lesion]	-0.08	-0.13 – -0.04	<**0.001**
blocked [unblocked] × delta	-0.09	-0.14 – -0.04	<**0.001**
group [lesion] × delta	0	-0.05 – 0.05	0.88
(blocked [unblocked] × group [lesion]) × delta	0.11	0.04 – 0.18	**0.003**
Observations	4153		
R2 / R2 adjusted	0.621 / 0.621		

Table 4.7: Linear regression model results Eq. (3.46) for Approach 1

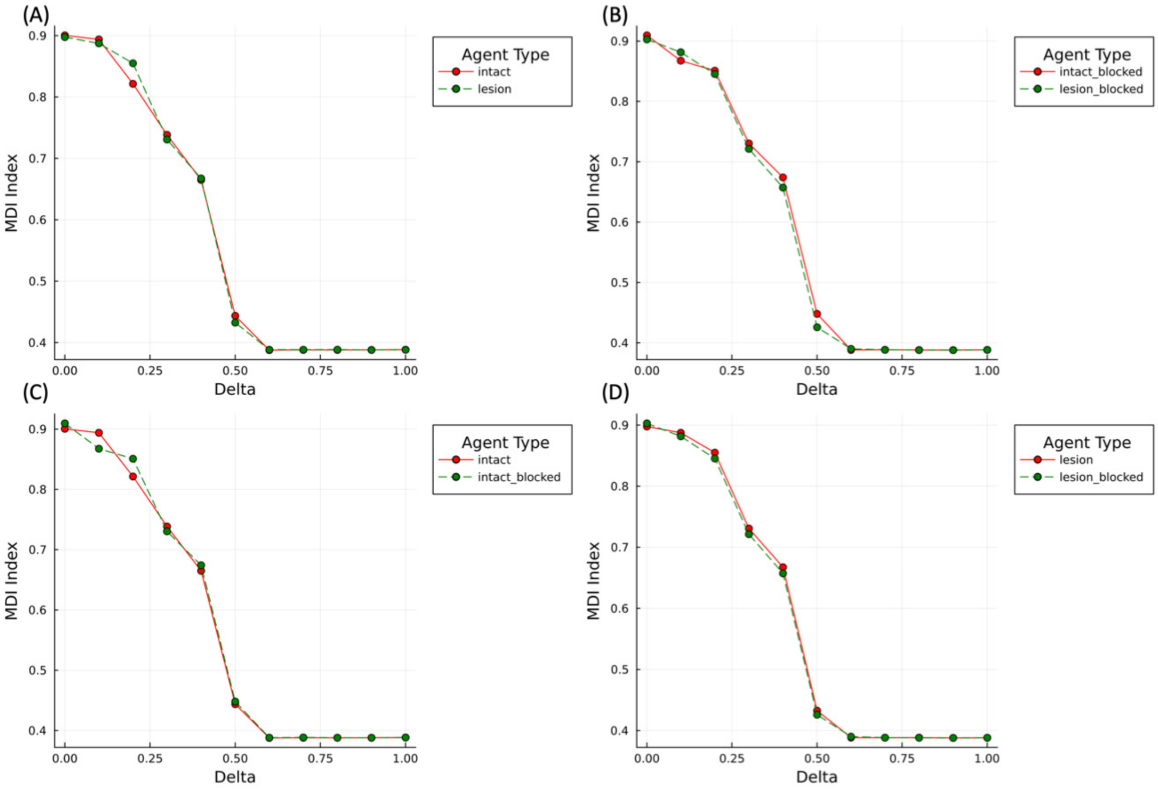


Figure 4.4: The MDI vs delta for **Approach 3**. **Panel (A):** The MDI vs delta for intact agents vs lesion agents. **Panel (B):** The MDI vs delta for intact agents vs lesion agents when the input is a blocked presentation. **Panel (C):** The MDI vs delta for intact agents when the words are encoded in blocked vs unblocked fashion. **Panel (D):** The MDI vs delta for lesion agents when the words are encoded in blocked vs unblocked fashion

negative effect ($\beta = -0.02, p = 0.005$; Table 4.8) and ($\beta = -0.06, p < 0.001$; Table 4.9).

Unblocked list presentation has a significant positive impact on the REC ($\beta = 0.01, p = 0.022$; Table 4.8) and ($\beta = 0.02, p = 0.023$; Table 4.9) consistent with results seen in Approach 1.

Approach 3: Recollection through MTL

Based on Fig. 4.7, the REC value is only significantly impacted by delta ($\beta = 0.3, p < 0.001$; Table 4.10) and ($\beta = 0.29, p < 0.001$; Table 4.11) and shows consistency with

	Dependent variable (*REC*)		
Predictors	**Estimates**	**CI**	**p**
(Intercept)	0.25	0.24 – 0.26	<**0.001**
blocked [unblocked]	0.01	0.00 – 0.02	**0.022**
group [lesion]	-0.02	-0.03 – -0.00	**0.005**
delta	0.63	0.62 – 0.64	<**0.001**
blocked [unblocked] × group [lesion]	-0.03	-0.04 – -0.01	**0.001**
Observations	4307		
R2 / R2 adjusted	0.686 / 0.686		

Table 4.8: Linear regression model results Eq. (3.45) for Approach 2

	Dependent variable (*REC*)		
Predictors	**Estimates**	**CI**	**p**
(Intercept)	0.27	0.25 – 0.28	<**0.001**
blocked [unblocked]	0.02	0.00 – 0.05	**0.023**
group [lesion]	-0.06	-0.08 – -0.03	<**0.001**
delta	0.59	0.56 – 0.62	<**0.001**
blocked [unblocked] × group [lesion]	-0.05	-0.08 – -0.02	**0.001**
blocked [unblocked] × delta	-0.02	-0.06 – 0.01	0.202
group [lesion] × delta	0.08	0.04 – 0.11	<**0.001**
(blocked [unblocked] × group [lesion]) × delta	0.05	0.00 – 0.10	**0.037**
Observations	4307		
R2 / R2 adjusted	0.691 / 0.691		

Table 4.9: Linear regression model results Eq. (3.46) for Approach 2

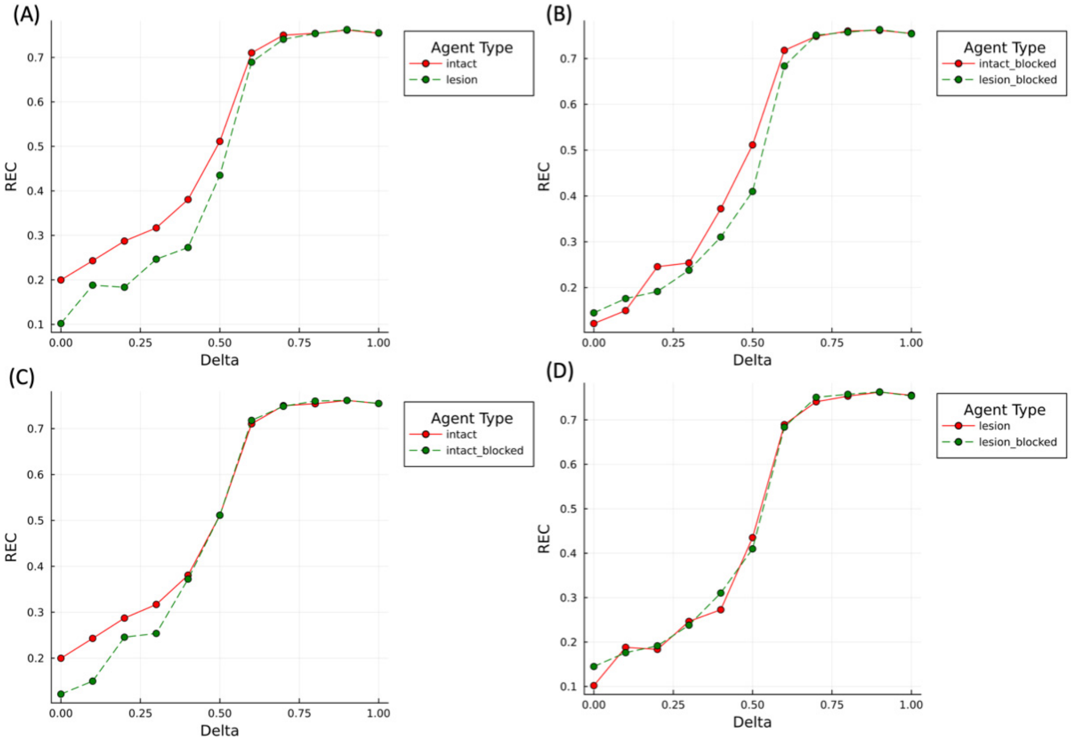


Figure 4.5: The REC vs delta for **Approach 1**. **Panel (A):** The REC vs delta for intact agents vs lesion agents. **Panel (B):** The REC vs delta for intact agents vs lesion agents when the input is a blocked presentation. **Panel (C):** The REC vs delta for intact agents when the words are encoded in blocked vs unblocked fashion. **Panel (D):** The REC vs delta for lesion agents when the words are encoded in blocked vs unblocked fashion

the results from Approach 1 in Section 4.2.2 and Approach 2 in Section 4.2.2.

	Dependent variable (*REC*)		
Predictors	**Estimates**	**CI**	**p**
(Intercept)	0.49	0.49 – 0.50	<**0.001**
blocked [unblocked]	0	-0.01 – 0.00	0.471
group [lesion]	0	-0.01 – 0.00	0.306
delta	0.3	0.29 – 0.30	<**0.001**
blocked [unblocked] × group [lesion]	0.01	-0.00 – 0.02	0.254
Observations	4397		
R2 / R2 adjusted	0.543 / 0.543		

Table 4.10: Linear regression model results Eq. (3.45) for Approach 3

	Dependent variable (*REC*)		
Predictors	**Estimates**	**CI**	**p**
(Intercept)	0.49	0.48 – 0.50	<**0.001**
blocked [unblocked]	-0.01	-0.02 – 0.01	0.437
group [lesion]	-0.01	-0.02 – 0.01	0.309
delta	0.29	0.28 – 0.31	<**0.001**
blocked [unblocked] × group [lesion]	0.01	-0.01 – 0.03	0.193
blocked [unblocked] × delta	0.01	-0.02 – 0.03	0.642
group [lesion] × delta	0.01	-0.02 – 0.03	0.577
(blocked [unblocked] × group [lesion]) × delta	-0.01	-0.05 – 0.02	0.413
Observations	4397		
R2 / R2 adjusted	0.543 / 0.542		

Table 4.11: Linear regression model results Eq. (3.46) for Approach 3

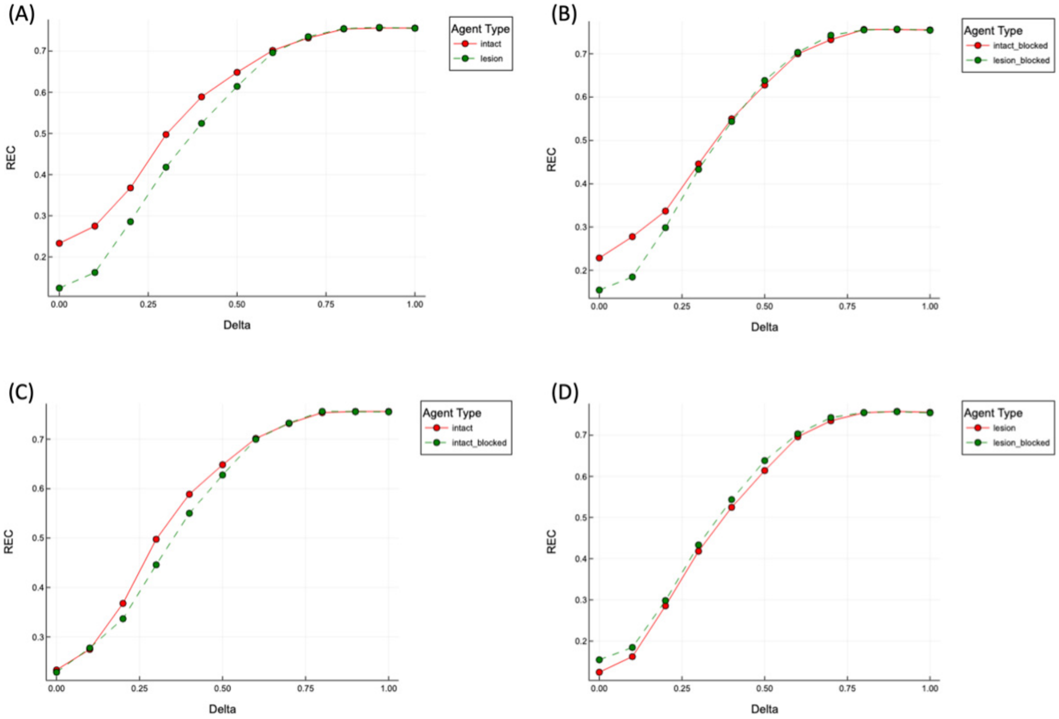


Figure 4.6: The REC vs delta for **Approach 2**. **Panel (A):** The REC vs delta for intact agents vs lesion agents. **Panel (B):** The REC vs delta for intact agents vs lesion agents when the input is a blocked presentation. **Panel (C):** The REC vs delta for intact agents when the words are encoded in blocked vs unblocked fashion. **Panel (D):** The REC vs delta for lesion agents when the words are encoded in blocked vs unblocked fashion

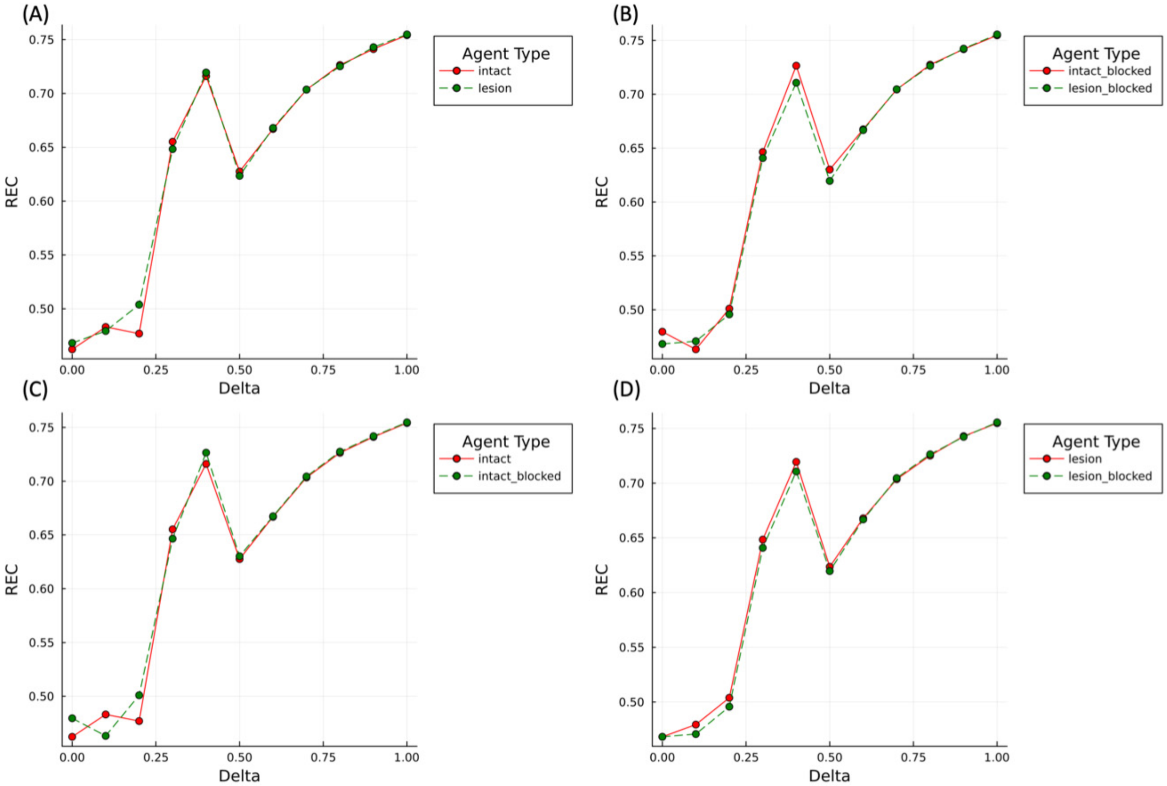


Figure 4.7: The REC vs delta for **Approach 3**. **Panel (A):** The REC vs delta for intact agents vs lesion agents. **Panel (B):** The REC vs delta for intact agents vs lesion agents when the input is a blocked presentation. **Panel (C):** The REC vs delta for intact agents when the words are encoded in blocked vs unblocked fashion. **Panel (D):** The REC vs delta for lesion agents when the words are encoded in blocked vs unblocked fashion

Chapter 5

Discussion and Conclusion

In the present study, we developed a computational model of prefrontal cortex-hippocampal interactions to perform the CVLT yes/no recognition task. We are the first study in the existing literature to measure the influence of prefrontal cortex and its lesions on verbal mnemonic discrimination performance through computational modeling approach. To accomplish this, we extend Becker and Lim's [10] model, which is the only model in literature that developed a prefrontal cortex-hippocampal model for the CVLT free recall task. Our extension involves replacing the binary representations of CVLT words with continuous word embeddings using ConceptNet [68], providing a more realistic representations of words. This representation helps us to capture the accurate semantic distances of words as interpreted in real world as opposed to simulated binary representations. Moreover, the use of the continuous representation of words facilitates the measurement of mnemonic discrimination performance in verbal recognition paradigm, employing a novel approach proposed by Leger et al. [40]. Our model incorporates the continuous Hopfield network as an autoassociator network for the hippocampus. We utilize the Ornstein-Uhlenbeck process as a context vector to represent continuous time which enables the representation of the temporal sequencing observed in the CVLT task. Furthermore, this enables us to accurately represent the temporal associations between words in the CVLT task, a factor known to influence the task performance [26, 32]. The prefrontal cortex is designed as a reinforcement learning agent that can learn to execute tasks depending on strategic requirements. For the yes/no recognition task, we implement three distinct approaches. The first one relies on full recollection which is based on previous research indicating that recognition performance relies on strategic retrieval of memories same as recall tasks [43, 85, 32]. In the second and third approaches we look at the individual contributions of the prefrontal cortex and the the hippocampus respectively. This will enable us to understand the influence of individual components on the mnemonic discrimination performance.

Mnemonic Discrimination Performance: The results of the mnemonic discrimination performance indicate that it is enhanced due to full recollection. Furthermore, the lesions in the prefrontal cortex impair the mnemonic discrimination performance and overall recognition performance. These findings support the view that strategic retrieval of memories plays an important role in mnemonic discrimination and that intact prefrontal cortex is required for the enhanced performance [39, 80, 29]. However, we observe that the overall recognition performance worsens due to recollection indicating a trade-off between mnemonic discrimination and overall recognition performance. This may arise due to a potential limitation in employing the recollection process of recall in recognition as discussed in Yonelinas et al. [85], given that both these processes may require different types of responses. While recall involves producing the studied words from memory, recognition process involves judging a given cue as old/new. These differences in test conditions could influence how individuals engage in the recollection process. For example, in our model, recollection involves the Hopfield network converging to one of the patterns stored in the memory. This approach is effective for recall, since the recall process simply requires production of words from memory and convergence of the Hopfield network will assist in this process. However, consider a scenario where a novel word from the recognition list undergoes this convergence. As the model always converges to the nearest stored pattern, it may fail to recognize the novel word as new. This phenomenon is also similar to *hallucinations* in large language models, in which the model generates a response that is nonsensical or unfaithful to the provided source content [27]. Understanding the nuances in retrieval strategies for recognition to prevent these false alarms can also guide us towards a better understanding of mitigating hallucinations in language models. Additionally, future studies can also look at incorporating confidence judgements to understand whether the recognition is based on recollection or is merely based on a sense of familiarity [76, 85]. This addition will provide a nuanced understanding of the cognitive processes involved in recognition tasks.

Another limitation in our study is the absence of list B recall of CVLT. The list B words in CVLT are a set of 16 words divided into 4 categories such that 2 categories overlap with List A, and the other 2 categories are distinct. The list B recall process acts as interference to the yes/no recognition task. Examining the effect of interference on the mnemonic discrimination performance could be an interesting potential research, especially considering previous research that have indicated the necessity of an intact executive control system to resolve interference during memory tasks [3, 66, 30]. These executive control systems, particularly the VLPFC, is hypothesized to mediate distinctive encoding in the hippocampus and facilitate interference resolution to support mnemonic discrimination. To explore this, future studies can incorporate a pattern separation mechanism to enable distinctive encoding in the hippocampus which is known to be facilitated by the Dentate Gyrus [21, 3].

Free Recall Performance: By extending the original Becker and Lim [23] model, we were able to replicate the effects of prefrontal cortex lesions on the free recall performance. This marks a significant contribution in understanding free recall paradigms using the continuous word embeddings. However, one challenging aspect of using the continuous Hopfield network as compared to the binary autoassociator network is its exponentially high storage capacity. Due to the high capacity and rapid learning of this Hopfield network, we were not able to replicate the dramatic learning slopes as reported in Becker and Lim [10] results. For example, in our model the recall capacity was approximately around 12 out of 16 words in the first trial, in contrast to Becker and Lim's model [10] which recalled only around 4 to 6 words out of 16. To address this discrepancy, we introduced an attention mechanism in the model. This mechanism is the phenomena observed during free recall tasks where participants often attend more to the initial words presented in the task, with attention decreasing towards the middle and end of the list [32]. Although this led to a learning curve, reducing the model's learning capacity in the first trial from 12 to 10, we still could not replicate the steep learning curves demonstrated in the original results. As a result, the high storage capacity of the continuous Hopfield network poses a limitation in our study. However, in real world data for CVLT, the learning curves differ for each participant. Thus, further studies can enhance the free recall algorithm such that these heterogeneous learning curves are captured.

In addition, recent study by Snow [105] has questioned the biological plausibility of the modern Hopfield network, indicating that the softmax computation is not biologically plausible in the brain. Recent study by Krotov et. al [106] proposed a biologically plausible modern Hopfield network but this network also has high storage capacity and so may not address the high storage capacity limitation of the Hopfield network.

We are the first study to propose a computational framework that outlines the role of the prefrontal cortex in a verbal mnemonic discrimination paradigm. We extend the existing model of free recall by Becker and Lim [10] to measure the mnemonic discrimination performance in the verbal yes/no recognition test. The results highlight the importance of an intact prefrontal cortex in the mnemonic discrimination and overall recognition performance. Our model can be a foundational framework that can be utilized to examine the role of executive functions like the prefrontal cortex in the verbal mnemonic discrimination paradigms. Future work should examine the use of confidence judgments during CVLT yes/no recognition to understand the degree to which this function depends on recollection vs. familiarity. Future studies should also capture the effects of interference from List B on yes/no recognition performance (including mnemonic discrimination). It would also be of great interest to understand the effects of non-local thresholding operations in the modern Hopfield network, as well as addition of pattern separating preprocessors (to model the dentate gyrus) in order to enhance the biological plausibility of our model. Finally, the predictions made by the model in the present study should be tested using real-world data in healthy controls and patients with frontal lesions.

Bibliography

[1] California Verbal Learning Test | Second Edition.

[2] ConceptNet.

[3] Tarek Amer and Lila Davachi. Extra-hippocampal contributions to pattern separation. *eLife*, 2023.

[4] John R. Anderson and Gordon H. Bower. Recognition and retrieval processes in free recall. *Psychological Review*, 79:97–123, 1972. Place: US Publisher: American Psychological Association.

[5] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web*, Lecture Notes in Computer Science, pages 722–735, Berlin, Heidelberg, 2007. Springer.

[6] Arnold Bakker, C. Brock Kirwan, Michael Miller, and Craig E. L. Stark. Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science (New York, N.Y.)*, 319(5870):1640–1642, March 2008.

[7] JULIANA V. BALDO, DEAN DELIS, JOEL KRAMER, and ARTHUR P. SHIMAMURA. Memory performance on the California Verbal Learning Test–II: Findings from patients with focal frontal lesions. *Journal of the International Neuropsychological Society*, 2002.

[8] Peter J. Bayley, John T. Wixted, Ramona O. Hopkins, and Larry R. Squire. Yes/No Recognition, Forced-choice Recognition, and the Human Hippocampus. *Journal of Cognitive Neuroscience*, 20:505–512, 2008.

[9] Jessica Bean. Rey Auditory Verbal Learning Test, Rey AVLT. In Jeffrey S. Kreutzer, John DeLuca, and Bruce Caplan, editors, *Encyclopedia of Clinical Neuropsychology*, pages 2174–2175. Springer, New York, NY, 2011.

[10] Suzanna Becker and Jean Lim. A computational model of prefrontal control in free recall: strategic memory use in the California Verbal Learning Task. *Journal of Cognitive Neuroscience*, 15(6):821–832, August 2003.

[11] Francis Bond and Ryan Foster. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[12] Jim Breen. JMdict: a Japanese-Multilingual Dictionary. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 65–72, Geneva, Switzerland, August 2004. COLING.

[13] Spyridon Chavlis and Panayiota Poirazi. Pattern separation in the hippocampus through the eyes of computational modeling. *Wiley*, 2017.

[14] C. Constantinidis, M. N. Franowicz, and P. S. Goldman-Rakic. Coding specificity in cortical microcircuits: a multiple-electrode analysis of primate prefrontal cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 21(10):3646–3655, May 2001.

[15] Patrick S R Davidson, Angela K Troyer, and Morris Moscovitch. Frontal lobe contributions to recognition and recall: Linking basic research with clinical evaluation and remediation. *J Int Neropsychol Soc.*, 2006.

[16] Colin A. Depp, Brent T. Mausbach, Alexandrea L. Harmell, Gauri N. Savla, Christopher R. Bowie, Philip D. Harvey, and Thomas L. Patterson. Meta-analysis of the association between cognitive abilities and everyday functioning in bipolar disorder. *Bipolar Disorders*, 14(3):217–226, May 2012.

[17] John Duncan. An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, 2(11):820–829, November 2001. Number: 11 Publisher: Nature Publishing Group.

[18] H. Eichenbaum, N. Fortin, M. Sauvage, R.J. Robitsek, and A. Farovik. An animal model of amnesia that uses Receiver Operating Characteristics (ROC) analysis to distinguish recollection from familiarity deficits in recognition memory. *Neuropsychologia*, 48(8):2281–2289, July 2010.

[19] Charles Elkan and Russell Greiner. Building large knowledge-based systems: Representation and inference in the cyc project: D.B. Lenat and R.V. Guha. *Artificial Intelligence*, 61(1):41–52, May 1993.

[20] C. D. Frith. The role of dorsolateral prefrontal cortex in the selection of action as revealed by functional imaging. *Control of cognitive processes*, 18:544–565, 2000. Publisher: The MIT Press Cambridge.

[21] Mark A. Gluck and Catherine E. Myers. *Gateway to Memory: An Introduction to Neural Network Modeling of the Hippocampus in Learning and Memory.* MIT Press, January 2001.

[22] M. F. Green, R. S. Kern, D. L. Braff, and J. Mintz. Neurocognitive deficits and functional outcome in schizophrenia: are we measuring the "right stuff"? *Schizophrenia Bulletin*, 26(1):119–136, 2000.

[23] Tae Hyon Ha, Ji Sun Kim, Jae Seung Chang, Sung Hee Oh, Ju Young Her, Hyun Sang Cho, Tae Sung Park, Soon Young Shin, and Kyooseob Ha. Verbal and Visual Memory Impairments in Bipolar I and II Disorder. *Psychiatry Investigation*, 9(4):339–346, December 2012.

[24] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2):245–258, July 2017.

[25] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, April 1982. Publisher: Proceedings of the National Academy of Sciences.

[26] Marc W. Howard and Michael J. Kahana. A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, 46(3):269–299, June 2002.

[27] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions, November 2023. arXiv:2311.05232 [cs].

[28] Sungmin Hwang, Enrico Lanza, Giorgio Parisi, Jacopo Rocchi, Giancarlo Ruocco, and Francesco Zamponi. On the Number of Limit Cycles in Diluted Neural Networks. *Journal of Statistical Physics*, 181(6):2304–2321, December 2020.

[29] Sarah A Johnson, Sabrina Zequeira, Sean M. Turner, Andrew P Maurer, Jennifer L Bizon, and Sara N Burke. Rodent mnemonic similarity task performance requires the prefrontal cortex. *Wiley*.

[30] John Jonides, Edward E. Smith, Christy Marshuetz, Robert A. Koeppe, and Patricia A. Reuter-Lorenz. Inhibition in verbal working memory revealed by brain activation. *Proceedings of the National Academy of Sciences of the United States of America*, 95(14):8410–8413, July 1998.

[31] Michael J. Kahana. Associative retrieval processes in free recall. *Memory & Cognition*, 24(1):103–109, January 1996.

[32] Michael J. Kahana. Computational Models of Memory Search. *Annual Review of Psychology*, 71(1):107–138, 2020. _eprint: https://doi.org/10.1146/annurev-psych-010418-103358.

[33] Eamonn Kennedy, Shashank Vadlamani, Hannah M. Lindsey, Pui-Wa Lei, Mary Jo-Pugh, Maheen Adamson, Martin Alda, Silvia Alonso-Lana, Sonia Ambrogi, Tim J. Anderson, Celso Arango, Robert F. Asarnow, Mihai Avram, Rosa Ayesa-Arriola, Talin Babikian, Nerisa Banaj, Laura J. Bird, Stefan Borgwardt, Amy Brodtmann, Katharina Brosch, Karen Caeyenberghs, Vince D. Calhoun, Nancy D. Chiaravalloti, David X. Cifu, Benedicto Crespo-Facorro, John C. Dalrymple-Alford, Kristen Dams-O'Connor, Udo Dannlowski, David Darby, Nicholas Davenport, John DeLuca, Covadonga M. Diaz-Caneja, Seth G. Disner, Ekaterina Dobryakova, Stefan Ehrlich, Carrie Esopenko, Fabio Ferrarelli, Lea E. Frank, Carol Franz, Paola Fuentes-Claramonte, Helen Genova, Christopher C. Giza, Janik Goltermann, Dominik Grotegerd, Marius Gruber, Alfonso Gutierrez-Zotes, Minji Ha, Jan Haavik, Charles Hinkin, Kristen R. Hoskinson, Daniela Hubl, Andrei Irimia, Andreas Jansen, Michael Kaess, Xiaojian Kang, Kimbra Kenney, Barbora Keřková, Mohamed Salah Khlif, Minah Kim, Jochen Kindler, Tilo Kircher, Karolina Knížková, Knut K. Kolskår, Denise Krch, William S. Kremen, Taylor Kuhn, Veena Kumari, Jun Soo Kwon, Roberto Langella, Sarah Laskowitz, Jungha Lee, Jean Lengenfelder, Spencer W. Liebel, Victoria Liou-Johnson, Sara M. Lippa, Marianne Løvstad, Astri Lundervold, Cassandra Marotta, Craig A. Marquardt, Paulo Mattos, Ahmad Mayeli, Carrie R. McDonald, Susanne Meinert, Tracy R. Melzer, Jessica Merchán-Naranjo, Chantal Michel, Rajendra A. Morey, Benson Mwangi, Daniel J. Myall, Igor Nenadić, Mary R. Newsome, Abraham Nunes, Terence O'Brien, Viola Oertel, John Ollinger, Alexander Olsen, Victor Ortiz García de la Foz, Mustafa Ozmen, Heath Pardoe, Marise Parent, Fabrizio Piras, Federica Piras, Edith Pomarol-Clotet, Jonathan Repple, Geneviève Richard, Jonathan Rodriguez, Mabel Rodriguez, Kelly Rootes-Murdy, Jared Rowland, Nicholas P. Ryan, Raymond Salvador, Anne-Marthe Sanders, Andre Schmidt, Jair C. Soares, Gianfranco Spalleta, Filip Španiel, Alena Stasenko, Frederike Stein, Benjamin Straube, April Thames, Florian Thomas-Odenthal, Sophia I. Thomopoulos, Erin Tone, Ivan Torres, Maya Troyanskaya, Jessica A. Turner, Kristine M. Ulrichsen, Guillermo Umpierrez, Elisabet Vilella, Lucy Vivash, William C. Walker, Emilio Werden, Lars T. Westlye, Krista Wild, Adrian Wroblewski, Mon-Ju Wu, Glenn R. Wylie, Lakshmi N. Yatham, Giovana B. Zunta-Soares, Paul M. Thompson, David F. Tate, Frank G. Hillary, Emily L. Dennis, and Elisabeth A. Wilde. Bridging Big Data: Procedures for Combining Non-equivalent Cognitive Measures from the ENIGMA Consortium, January 2023. Pages: 2023.01.16.524331 Section: New Results.

[34] Danielle R. King, Marianne de Chastelaine, Rachael L. Elward, Tracy H. Wang, and Michael D. Rugg. Recollection-Related Increases in Functional Connectivity Predict Individual Differences in Memory Accuracy. *Journal of Neuroscience*, 35(4):1763–1772, January 2015. Publisher: Society for Neuroscience Section: Articles.

[35] R. M. Klein. Inhibition of return. *Trends in Cognitive Sciences*, 4(4):138–147, April 2000.

[36] Jenna L. Klippenstein, Shauna M. Stark, Craig E. L. Stark, and Ilana J. Bennett. Neural substrates of mnemonic discrimination: A whole-brain fMRI investigation. *Brain and Behavior*, 2020.

[37] Teuvo Kohonen. *Content-Addressable Memories*, volume 1 of *Springer Series in Information Sciences*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1980.

[38] Dmitry Krotov and John J. Hopfield. Dense Associative Memory for Pattern Recognition, September 2016. arXiv:1606.01164 [cond-mat, q-bio, stat].

[39] Claire Lauzon. *THE ROLE OF THE VENTROMEDIAL PREFRONTAL CORTEX IN MNEMONIC DISCRIMINATION AND GENERALIZATION*. PhD thesis, YORK UNIVERSITY, Toronto, 2022.

[40] Simon Leger, Christian Guinard, Selena Singh, Suzanna Becker, Jasmyn E A Cunningham, Martin Alda, Aaron J Newman, Thomas Trappenberg, and Abraham Nunes. A new measure of mnemonic discrimination applicable to recognition memory tests with continuous variation in novel stimulus interference, Nov 2023.

[41] Alexander M. P., Stuss D. T., and Fansabedian N. California Verbal Learning Test: performance by patients with focal frontal and non-frontal lesions. 2003.

[42] Tom Macpherson, Anne Churchland, Terry Sejnowski, James DiCarlo, Yukiyasu Kamitani, Hidehiko Takahashi, and Takatoshi Hikida. Natural and Artificial Intelligence: A brief introduction to the interplay between AI and neuroscience research. *Neural Networks: The Official Journal of the International Neural Network Society*, 144:603–613, December 2021.

[43] George Mandler. Recognizing: The judgment of previous occurrence. *Psychological Review*, 87(3):252–271, 1980. Place: US Publisher: American Psychological Association.

[44] Joseph R. Manns, Ramona O. Hopkins, Jonathan M. Reed, Erin G. Kitchener, and Larry R. Squire. Recognition Memory and the Human Hippocampus. *Neuron*, 37(1):171–180, January 2003. Publisher: Elsevier.

[45] D. Marr. Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 262(841):23–81, July 1971.

[46] B. L. McNaughton and R. G. M. Morris. Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, 10(10):408–415, January 1987.

[47] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality, October 2013. arXiv:1310.4546 [cs, stat].

[48] Morris Moscovitch and Gordon Winocur. The Frontal Cortex and Working with Memory. In Donald T. Stuss and Robert T. Knight, editors, *Principles of Frontal Lobe Function*, page 0. Oxford University Press, August 2002.

[49] Catherine E. Myers, Keria Bermudez-Hernandez, and Helen E. Scharfman. The influence of ectopic migration of granule cells into the hilus on dentate gyrus-CA3 function. *PloS One*, 8(6):e68208, 2013.

[50] Catherine E. Myers and Helen E. Scharfman. A role for hilar cells in pattern separation in the dentate gyrus: a computational approach. *Hippocampus*, 19(4):321–337, April 2009.

[51] Catherine E. Myers and Helen E. Scharfman. Pattern separation in the dentate gyrus: a role for the CA3 backprojection. *Hippocampus*, 21(11):1190–1215, November 2011.

[52] D. A. Nathaniel-James and C. D. Frith. The Role of the Dorsolateral Prefrontal Cortex: Evidence from the Effects of Contextual Constraint in a Sentence Completion Task. *NeuroImage*, 16(4):1094–1102, August 2002.

[53] Rapeechai Navawongse and Howard Eichenbaum. Distinct pathways for rule-based retrieval and spatial mapping of memory representations in hippocampal neurons. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 33(3):1002–1013, January 2013.

[54] Yael Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154, June 2009.

[55] Randall O' Reilly, Kenneth Norman, and James McClelland. A Hippocampal Model of Recognition Memory. In *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997.

[56] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[57] Alison R. Preston and Howard Eichenbaum. Interplay of hippocampus and prefrontal cortex in memory. *Current biology : CB*, 23(17):R764–R773, September 2013.

[58] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield Networks is All You Need, April 2021. arXiv:2008.02217 [cs, stat].

[59] RA Rescorla and Allan Wagner. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory*, volume Vol. 2. January 1972. Journal Abbreviation: Classical Conditioning II: Current Research and Theory.

[60] Edmund T. Rolls. A quantitative theory of the functions of the hippocampal CA3 network in memory. *Frontiers in Cellular Neuroscience*, 7:98, June 2013.

[61] Jacob Russin, Randall C. O'Reilly, and Yoshua Bengio. Deep Learning Needs a Prefrontal Cortex. 2020.

[62] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science (New York, N.Y.)*, 275(5306):1593–1599, March 1997.

[63] Wolfram Schultz. Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, 18(1):23–32, March 2016.

[64] Per B. Sederberg, Marc W. Howard, and Michael J. Kahana. A context-based theory of recency and contiguity in free recall. *Psychological Review*, 115(4):893–912, October 2008.

[65] Carol A. Seger and Earl K. Miller. Category Learning in the Brain. *Annual review of neuroscience*, 33:203–219, 2010.

[66] Arthur P. Shimamura, Paul J. Jurica, Jennifer A. Mangels, Felicia B. Gershberg, and Robert T. Knight. Susceptibility to memory interference effects following frontal lobe damage: Findings from tests of paired-associate learning. *Journal of Cognitive Neuroscience*, 7(2):144–152, 1995. Place: US Publisher: MIT Press.

[67] Push Singh. The Public Acquisition of Commonsense Knowledge. 2002.

[68] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *CoRR*, abs/1612.03975, 2016. arXiv: 1612.03975.

[69] Shauna M. Stark, C. Brock Kirwan, and Craig E.L. Stark. Mnemonic Similarity Task: A Tool for Assessing Hippocampal Integrity. *Trends in cognitive sciences*, 23(11):938–951, November 2019.

[70] Donald T. Stuss, Michael P. Alexander, Carole L. Palumbo, Leslie Buckle, Lisa Sayer, and Janice Pogue. Organizational strategies with unilateral or bilateral frontal lobe injury in word learning tasks. *Neuropsychology*, 8(3):355–373, 1994. Place: US Publisher: American Psychological Association.

[71] Tomiki Sumiyoshi. Verbal memory. *Handbook of Experimental Pharmacology*, 228:237–247, 2015.

[72] Richard S. Sutton and Andrew G. Barto. Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88(2):135–170, 1981. Place: US Publisher: American Psychological Association.

[73] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* A Bradford Book, Cambridge, MA, USA, October 2018.

[74] Jack E. Taylor, Alistair Beith, and Sara C. Sereno. LexOPS: An R package and user interface for the controlled generation of word stimuli. *Behavior Research Methods*, 52(6):2372–2382, December 2020.

[75] I. J. Torres, C. M. DeFreitas, V. G. DeFreitas, D. J. Bond, M. Kunz, W. G. Honer, R. W. Lam, and L. N. Yatham. Relationship between cognitive functioning and 6-month clinical and functional outcome in patients with first manic episode bipolar I disorder. *Psychological Medicine*, 41(5):971–982, May 2011.

[76] Endel Tulving. Memory and consciousness. *Canadian Psychology / Psychologie canadienne*, 26(1):1–12, 1985. Place: Canada Publisher: Canadian Psychological Association.

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. arXiv:1706.03762 [cs].

[78] Luis von Ahn, Mihir Kedia, and Manuel Blum. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 75–78, New York, NY, USA, April 2006. Association for Computing Machinery.

[79] Peter E. Wais, Sahar Jahanikia, Daniel Steiner, Craig E. L. Stark, and Adam Gazzaley. Retrieval of high-fidelity memory arises from distributed cortical networks.

[80] Peter E. Wais, Olivia Montgomery, Craig E. L. Stark, and Adam Gazzaley. Evidence of a Causal Role for mid- Ventrolateral Prefrontal Cortex Based Functional Networks in Retrieving High-Fidelity Memory. *Scientific Reports*, 2018.

[81] Jane X. Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21(6):860–868, June 2018. Number: 6 Publisher: Nature Publishing Group.

[82] James C. R. Whittington, Joseph Warren, and Timothy E. J. Behrens. Relating transformers to models and neural representations of the hippocampal formation, March 2022. arXiv:2112.04035 [cs, q-bio].

[83] Michael Widrich, Bernhard Schäfl, Hubert Ramsauer, Milena Pavlović, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, and Günter Klambauer. Modern Hopfield Networks and Attention for Immune Repertoire Classification, July 2020. arXiv:2007.13505 [cs, q-bio, stat].

[84] Michael A. Yassa, Aaron T. Mattfeld, Shauna M. Stark, and Craig E. L. Stark. Age-related memory deficits linked to circuit-specific disruptions in the hippocampus. *Proceedings of the National Academy of Sciences of the United States of America*, 108(21):8873–8878, May 2011.

[85] Andrew P Yonelinas. The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, 46(3):441–517, April 2002.

[86] Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, James DiCarlo, Surya Ganguli, Jeff Hawkins, Konrad Körding, Alexei Koulakov, Yann LeCun, Timothy Lillicrap, Adam Marblestone, Bruno Olshausen, Alexandre Pouget, Cristina Savin, Terrence Sejnowski, Eero Simoncelli, Sara Solla, David Sussillo, Andreas S. Tolias, and Doris Tsao. Catalyzing next-generation Artificial Intelligence through NeuroAI. *Nature Communications*, 14(1):1597, March 2023. Number: 1 Publisher: Nature Publishing Group.

Appendix A

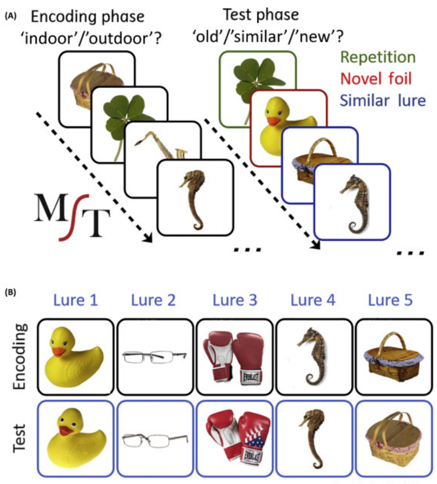
Mnemonic Similarity Task

A.1 Mnemonic Discrimination - Mnemonic Similarity Task

The gold standard task for assessing the mnemonic discrimination ability is the object recognition memory task called the Mnemonic Similarity Task (MST) developed by Stark et al. [69]. This task is used clinically to study the hippocampal dysfunction in humans with neurological diseases like depression and schizophrenia.

The traditional version of MST has two phases. The initial phase is a study phase where the participants are shown images of everyday objects and are asked to judge if they are *indoor/outdoor*. This is immediately followed by a recognition test phase where the participants should identify if each image is *new*, *old* or *similar*. Fig. A.1 gives an example of the images shown during the study phase and test phase [69]. The images in the test phase are equally divided into *targets* (images that are exact repetitions of study phase), *foils* (new images) and *lures* (images that are perpetually similar to the ones in study phase but are not identical). The main goal of MST is to analyze the ability to discriminate the lures and flag them as *similar* instead of *old* which measures the degree of pattern separation in the brain.

To calculate the performance of MST, Lure Discrimination Index (LDI) and Recognition Memory performance (REC) are calculated. The LDI is the difference between probability of a lure image correctly being identified as *similar* and the probability that a foil image is mistakenly identified as *similar*. The LDI performance relies a lot on pattern separation and is critical when assessing hippocampal integrity [69]. The REC score is the difference between probability of an old image correctly being identified as *old* and the probability that a foil image is mistakenly identified as *old*. While LDI measures mnemonic discrimination, REC measures the overall recognition performance.



Trends in Cognitive Sciences

Figure A.1: Mnemonic Similarity Task. **Panel (A):** Left side are images in the encoding phase where they should be identified as *indoor/outdoor*. Images on the right are shown during the test phase where they should be classified as *old/new/similar*. **Panel (B):** The lures shown during encoding and test phase. (Source [69])

Appendix B

Importance of Assessing Mnemonic Discrimination on Verbal Memory

Below we outline the reasons why analyzing mnemonic discrimination in verbal memory is crucial and offers a distinct advantage compared to object recognition tasks.

1. Verbal memory, which relates to storage of language and vocabulary related information, plays a crucial role in influencing functional aspects like engagement in community activities, problem solving in social contexts and acquisition of new skills [71]. Moreover, patients with psychiatric disorders like bipolar disorder and schizophrenia often exhibit significant impairments in verbal memory. Consequently, examining verbal memory has always been a subject of interest in studies involving patients with neuropsychiatric disorders. Numerous meta-analysis and longitudinal studies have been conducted to ascertain the impact of verbal memory impairments on the day-to-day functioning of individuals with neuropsychiatric disorders [22, 75, 16]. All these studies consistently affirm the significant effect of verbal memory impairments on functional well-being of individuals with neuropsychiatric disorders. These factors emphasize a crucial need for exploring various aspects of verbal memory paradigms.

2. Given the importance of verbal memory in humans, memory tests like CVLT and RAVLT [9] have been used extensively to study verbal memory impairments in humans. However, these studies have always focused on verbal recall and recognition memory paradigms, overlooking the aspect of mnemonic discrimination. So, it is important to bridge this gap by exploring mnemonic discrimination in the context of verbal learning tasks, which has the potential to yield insightful revelations.

3. Since verbal memory tests are most commonly used in both research and clinical settings, they provide access to extensive and diverse clinical datasets. A recent study conducted by Enhancing Neuroimaging Genetics through Meta-Analysis (ENIGMA) consortium [33], involved data harmonization from 53 studies consisting of nearly 10,000 healthy and brain injured individuals. Also, the mnemonic discrimination paradigm has primarily been studied using the Mnemonic Similarity Task, limiting the amount of data available for its study. Therefore, assessing mnemonic discrimination within verbal learning tasks would provide us access with large ecologically relevant datasets for analysis of mnemonic discrimination across many conditions.

Appendix C

Power Calculations for Free Recall Results

u (F-statistic numerator degrees of freedom - number of coefficients in the model without the intercept)	4
v (F-statistic denominator degrees of freedom)	1995
f2 ($R^2/(1-R^2)$)	0.2536041
Sig.level (Significance level)	0.001
power	1

Table C.1: Power calculation for Model 1 (Eq. (3.42))

u (F-statistic numerator degrees of freedom - number of coefficients in the model without the intercept)	4
v (F-statistic denominator degrees of freedom)	1995
f2 ($R^2/(1-R^2)$)	0.2033694
Sig.level (Significance level)	0.001
power	1

Table C.2: Power calculation for Model 2 (Eq. (3.43))

Appendix D

Power Calculations for MDI

D.1 Approach 1: Full Recollection

The Table D.1 represent the Power Calculations for Eq. (3.44) for Approach 1

u (F-statistic numerator degrees of freedom - number of coefficients in the model without the intercept)	7
v (F-statistic denominator degrees of freedom)	4145
f2 ($R^2/(1-R^2)$)	0.1875074
Sig.level (Significance level)	0.001
power	1

Table D.1: Power calculation for Eq. (3.44) in Approach 1

D.2 Approach 2: Recollection through PFC

The Table D.2 represent the Power Calculations for Eq. (3.44) for Approach 2

u (F-statistic numerator degrees of freedom - number of coefficients in the model without the intercept)	7
v (F-statistic denominator degrees of freedom)	4299
f2 ($R^2/(1-R^2)$)	0.3241525
Sig.level (Significance level)	0.001
power	1

Table D.2: Power calculation for Eq. (3.44) in Approach 2

D.3 Approach 3: Recollection through MTL

The Table D.3 represent the Power Calculations for Eq. (3.44) for Approach 3

u (F-statistic numerator degrees of freedom - number of coefficients in the model without the intercept)	7
v (F-statistic denominator degrees of freedom)	4389
f2 ($R^2/(1-R^2)$)	2.699593
Sig.level (Significance level)	0.001
power	1

Table D.3: Power calculation for Eq. (3.44) in Approach 3

Appendix E

Power Calculations for REC

E.1 Approach 1: Full Recollection

The Tables E.1 and E.2 represent the Power Calculations for Eq. (3.45) and Eq. (3.46) respectively for Approach 1

u (F-statistic numerator degrees of freedom - number of coefficients in the model without the intercept)	4
v (F-statistic denominator degrees of freedom)	4148
f2 ($R^2/(1-R^2)$)	1.624672
Sig.level (Significance level)	0.001
power	1

Table E.1: Power calculation for Eq. (3.45) in Approach 1

u (F-statistic numerator degrees of freedom - number of coefficients in the model without the intercept)	7
v (F-statistic denominator degrees of freedom)	4145
f2 ($R^2/(1-R^2)$)	1.639916
Sig.level (Significance level)	0.001
power	1

Table E.2: Power calculation for Eq. (3.46) in Approach 1

E.2 Approach 2: Recollection through PFC

The Tables E.3 and E.4 represent the Power Calculations for Eq. (3.45) and Eq. (3.46) respectively for Approach 2

u (F-statistic numerator degrees of freedom - number of coefficients in the model without the intercept)	4
v (F-statistic denominator degrees of freedom)	4302
f2 ($R^2/(1-R^2)$)	2.184713
Sig.level (Significance level)	0.001
power	1

Table E.3: Power calculation for Eq. (3.45) in Approach 2

u (F-statistic numerator degrees of freedom - number of coefficients in the model without the intercept)	7
v (F-statistic denominator degrees of freedom)	4299
f2 ($R^2/(1-R^2)$)	2.237294
Sig.level (Significance level)	0.001
power	1

Table E.4: Power calculation for Eq. (3.46) in Approach 2

E.3 Approach 3: Recollection through MTL

The Tables E.5 and E.6 represent the Power Calculations for Eq. (3.45) and Eq. (3.46) respectively for Approach 3

u (F-statistic numerator degrees of freedom - number of coefficients in the model without the intercept)	4
v (F-statistic denominator degrees of freedom)	4392
f2 ($R^2/(1-R^2)$)	1.187705
Sig.level (Significance level)	0.001
power	1

Table E.5: Power calculation for Eq. (3.45) in Approach 3

u (F-statistic numerator degrees of freedom - number of coefficients in the model without the intercept)	7
v (F-statistic denominator degrees of freedom)	4389
f2 ($R^2/(1-R^2)$)	1.188184
Sig.level (Significance level)	0.001
power	1

Table E.6: Power calculation for Eq. (3.46) in Approach 3

Appendix F

Key Motivation to Build a Computational Model

Advancements in artificial intelligence (AI) and neuroscience have always been closely linked [42, 24, 86]. Major breakthroughs in AI have created neural networks based on the properties of cognitive systems. For example, deep reinforcement learning models are based on models originally used to understand classical and operant behavioral conditioning in animals [59, 72]. Moreover, developments in AI are also being utilized as a tool for neuroscience research to understand the functioning of the brain. For example, spin-glass models and autoassociative attractor networks have been used to understand memory processing in the hippocampal CA3 region [28, 25]. Autoencoders [21] and transformer models [82] have each also been proposed as models of overall hippocampal functioning. Similarly, AI models designed to perform reinforcement learning have helped neuroscientists provide an interpretation of how dopamine neurons mediate reward dependent learning in the brain [42, 62, 63].

F.1 Key Motivation to Build a Computational Model For Our Study

A key motivation for utilizing AI to build computational models of the brain is that this enables the simulation of specific brain functions and various experimental conditions that would be unethical to perform on humans. This is particularly important for research involving human neurological disorders involving language-based memory, since

1. Only humans possess language, thereby precluding its study using animal models.

2. The function of language-related microcircuits cannot be feasibly studied in vivo in large samples of patients and healthy controls.

Therefore, the use of computational modeling approaches to study brain functioning may help guide neuroscience research toward understanding the biological basis of cognition in health and disease.