Hippocampo-neocortical interaction as compressive retrieval-augmented generation

Eleanor Spens & Neil Burgess
Institute of Cognitive Neuroscience, University College London

Many cognitive functions involve interplay between episodic (hippocampal) and semantic (neocortical) systems, but the mechanisms are unclear. We present a computational model in which sequential experiences are encoded in hippocampus in compressed form and replayed to train a neocortical generative network. This network captures the gist of specific episodes and extracts statistical patterns that generalise to new situations, enabling efficient reconstruction of the past and prediction of the future. The two systems interact during recall and prediction, with the hippocampus retrieving relevant episodic information into working memory as a basis for generation using the 'general knowledge' of the neocortical network. We simulate this interaction as 'retrieval-augmented generation', with the addition of mechanisms to *compress* episodic memories into hippocampus and to *consolidate* them into neocortex. The model explains changes to memories over time, including schema-based distortions, and shows how recent episodic and semantic memory contribute to new problem solving.

1 Introduction

Despite the success of the theorised distinction between episodic (hippocampal) and semantic (neocortical) systems, and long-standing ideas regarding the consolidation of memories from hippocampus to neocortex (*1–4*), how both systems work together to support a wide variety of cognitive functions is unclear. These include recollection and problem solving, which rely on a variable mixture of specific episodic memories from the hippocampus *and* more general knowledge from the neocortex (*5, 6*). Here we lay out a framework for how this co-operation occurs, and make use of recent advances in machine learning to simulate how the resulting system performs these functions for comparison with human performance. This builds on many years of theoretical insights into the relationship between schemas and memory (*7–12*), and their potential correspondence to large language models (*13, 14*).

We hypothesise that:

1. A deep generative model in neocortex is trained to predict its own inputs, using memories of episodic experiences replayed from hippocampus (and during the experiences themselves). In this way it develops generalisable (semantic) knowledge, enabling it to generate episodes from an initial input.

2. The hippocampus stores episodic memories in a highly compressed form, as a minimal subset of information from which the generative model can reconstruct each episode, combining both general schemas and episode-unique features.

3. Hippocampus and neocortex work together in recall to provide episodic details and predictions from semantic knowledge given these details respectively. But with enough consolidation, neocortex can

approximately reconstruct memories without the hippocampus, allowing the redundant hippocampal trace to fade away.

4. Similarly, episodic contributions to planning and problem-solving correspond to the hippocampus retrieving relevant memories on which to base neocortical generation into working memory.

In summary, we propose that highly compressed hippocampal memories relevant to a task are loaded into working memory as context for a neocortical generative network, which in turn is trained on these memories. The resulting model develops generalisable semantic knowledge from specific events. Our model reflects the complex interplay between episodic and semantic memory; semantic memory is acquired by learning to reconstruct episodic memories, but also shapes the representation of these episodes, both in terms of their initial encoding and their gradual schematisation.

Here working memory refers to the temporary retention of a limited amount of information (*15*), which can include episodic memories retrieved from long-term memory (*16*). To be more precise, we use 'working memory' as shorthand for the episodic buffer, a temporary store thought to sit between the phonological loop / visuospatial sketchpad and long term memory (*16*), which can accommodate considerably more than the 4 or so items which can be attended to at a given moment (*16, 17*). This is the 'workspace' into which episodic memories relevant to a task are retrieved from the hippocampus, upon which the neocortical network operates.

We first consider simple narratives representing autobiographical episodes, and demonstrate i) their encoding in compressed form, ii) recall as retrieval-augmented generation, in which the generative network 'decompresses' the event, and iii) consolidation. We show that the model predicts changes to the content of memory over time, such as greater abstraction and the loss of specific detail, as seen in human memory (*18,19*). We also show that the compression aspect of the model enables efficient storage and reconstruction, suggesting how rich and complex events can be stored over a lifetime without exceeding limited hippocampal capacity. Next, we show how encoding and consolidation don't just result in loss of detail, but systematically distort memories towards expectations in our model, matching the experimental data (*20, 21*). Finally, we show how consolidation develops generalisable knowledge in the neocortical network (*22*), allowing problem-solving by combining this with relevant specifics in context, which are loaded into working memory from hippocampus or ongoing experience. Our model works with both linguistic and non-linguistic stimuli, and is thus applicable to sequential experience of many kinds.

Implementing the model via compressive retrieval augmented generation

Here we use transformer-based large language models (*23*) to simulate how neocortical networks combine with the hippocampus to extract semantic knowledge from sequential experiences, and to support memory and problem-solving.

Machine learning has shown how deep neural networks can act as task-general 'foundation models' (*24*). In particular, large language models (LLMs) demonstrate that complex sequential behaviours can develop as a byproduct of a simple 'next item prediction' task (*23, 25*), with distributed semantic representations slowly learned over time as in neocortex (*26*). Here we suggest how memory consolidation could contribute to the development of neural 'foundation models' in a similar way, with neocortex learning to reconstruct replayed memories via schemas (see (*27*)). Crucially, LLMs can memorise specific sequences as well as learning generalities (*28*), meaning that episodic and semantic information can be captured within a single network (once the former is fully consolidated). We model consolidation as the training of autoregressive sequence models, simulated using Mistral 7B (*29*) and GPT-2 (*23*), on sequences which represent replayed hippocampal memories. Consolidation can thus be thought of as teacher-student learning (see (*30*)). The objective during training is simply to predict the next item in sequences from the training data. Once trained,

the network can continue an input sequence, or generate a new sequence from scratch, by iteratively predicting the next item from the items so far (the 'prompt' or 'context').

We propose that the hippocampal 'memory bank' and neocortical generative network work together to enable problem-solving, with the hippocampus retrieving memories relevant to a new task into working memory as context for the generative network. Specifically, one can draw inspiration from retrieval-augmented generation (RAG; (*31*)), which involves prompting LLMs with relevant data retrieved from external memory, as a model of drawing on relevant memories to inform predictions about a new event (see also (*13*)). A typical task for RAG might be question answering based on a set of new facts which do not feature in the training data for the chosen LLM. To ask the system a question, first the most relevant items are found in the external memory (using some similarity metric). A 'prompt' is then constructed, in which the retrieved items are given to the LLM together with the question.

We suggest hippocampal-neocortical interaction may resemble RAG, but crucially with additional compression and consolidation mechanisms. Memories in the hippocampal 'retriever' are stored in a compact conceptual form, as in the LLM literature on context compression (e.g. (*32*)), and consolidation gradually trains the neocortical 'generator' on memories from the 'retriever'. The conceptual representations in hippocampus result from compressing experience in a way that is optimal for efficient reconstruction of memories, and all recall of hippocampal traces is retrieval-augmented generation, requiring the generative network to fill in the outline of a memory with its predictions.

The 'context' (or 'prompt') provided to an LLM can be thought of as its 'working memory' (*33*), with the limited context length of an LLM representing the limited capacity of biological working memory. Note that the neocortical model's outputs could be conditioned on perception rather than memory by the same mechanism, if the contents of working memory are new sequences rather than ones retrieved from the hippocampus.

Previous work

There is a long history of examining how short, medium, and long-term stores support encoding, consolidation, and recall across the 'lifespan' of a memory (e.g. (*34*)). Crucially, semantic memory is thought to be extracted from episodic memory through consolidation (*2, 3, 35*), learning statistical relationships from experience to aid prediction and thus survival (*36*). Both semantic and episodic memory can interact within working memory when reconstructing what happened in the past or imagining what will happen in the future (*37*), with episodic memory contributing specifics and semantic memory contributing generalisable knowledge (*38*). The conception of memory as (re)constructed rather than veridical (*20*) highlights the potential of recently developed deep generative networks to shed new light on these questions.

The idea that memories are rapidly encoded by the hippocampus and then replayed over the course of systems consolidation to train a generative or predictive model of the world in neocortex has received much recent interest. The combined system is thought to support multiple cognitive functions including episodic memory, semantic memory, imagination, and inference (*27, 30, 39–41*). This provides a mechanistic account of the theory that episodic memories are reconstructions that are influenced by our beliefs, i.e. that recall involves 'predicting' the past (*42–44*). However previous work has focused on static patterns, whereas episodes are sequential, and arguably the most crucial function of semantic knowledge involves prediction of what will come next.

Memory helps to predict the future, not just recall the past (*36*). Planning often requires anticipating the outcome of a possible action, e.g. the sequence of future states and rewards that would follow it. The mental simulation of future states to aid decision-making is known as 'model-based' planning (*45, 46*), which improves with consolidation (*47*). We suggest that the generative network trained on memories can simulate events to support behaviour, such as when deciding between possible actions, or when inferring the next state given the sequence so far. (See also (*48*).)

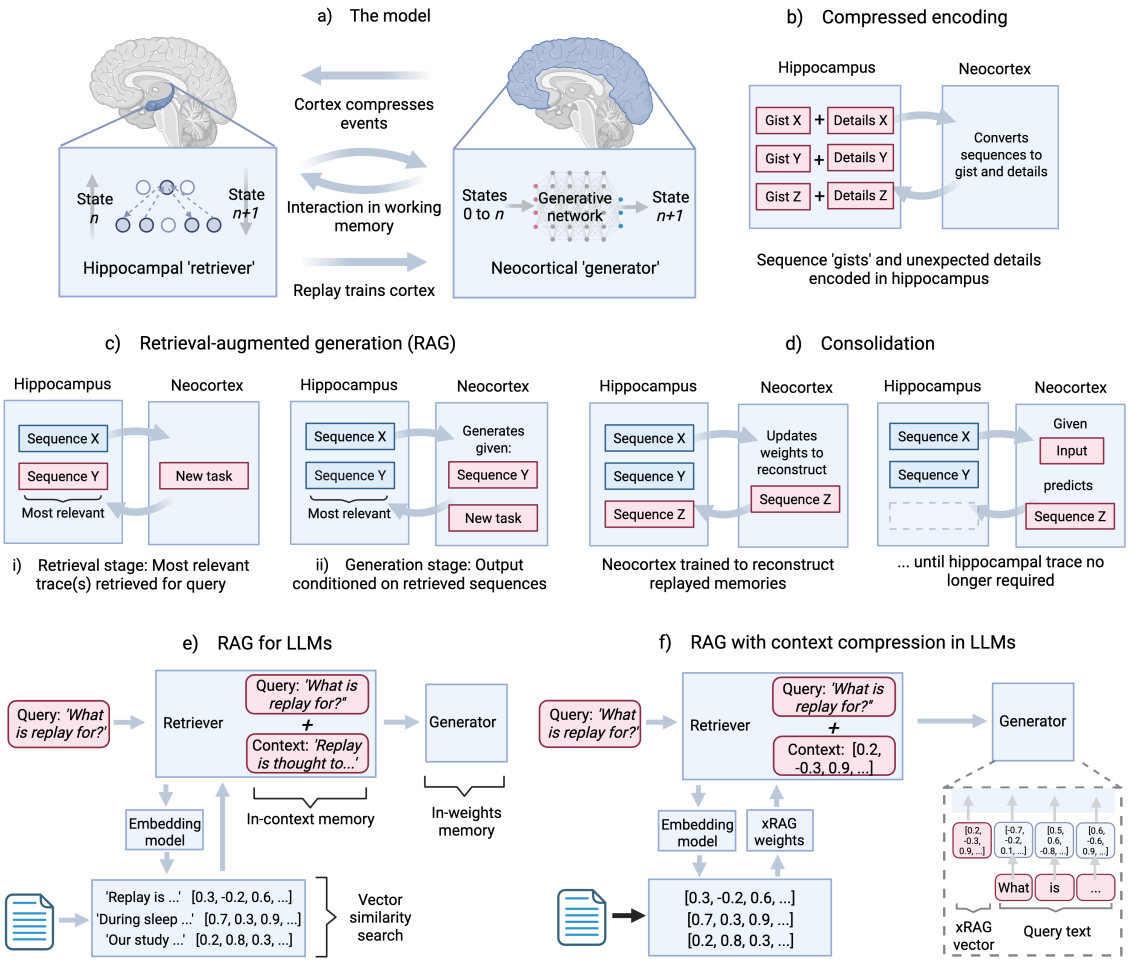

Figure 1: Summary of model. a) Episodic memories are first encoded in the hippocampus, then replayed during rest to train a generative network (simulated by an LLM) which thereby acquires semantic knowledge. The hippocampus ('retriever') and neocortex ('generator') can interact to play complementary roles during problem solving, resembling 'retrieval-augmented generation' (RAG; (*31*)). Relevant hippocampal traces are retrieved into working memory and then used to 'prompt' the neocortical network. b) Sequences are encoded in the hippocampus in compressed form. Specifically, a vector from which the sequence can be approximately decoded is stored for each event, together with the subsequences that were most surprising given this compressed version of the event. c) The neocortical and hippocampal networks can work together to recall memories and solve problems via RAG. d) Memories are replayed from the hippocampus during rest, and the generative network learns to capture their statistics through prediction error minimisation. Once a sequence is consolidated, the generative network can retrieve the rest of a sequence given its start, without requiring the hippocampal trace. e) RAG for LLMs. When the neocortical generative network ('generator') is prompted with a query, related information is retrieved from hippocampal memory ('retriever'). The retrieval process finds stored information with similar vectors (produced by a 'text embedding model') to the query. A combined prompt is then created from the query and the retrieved information, improving the output from the generator. f) RAG with context compression using xRAG (*32*). A small network, trained with the LLM, maps vectors from the text embedding model into a single token representation. This allows the retrieved information to be represented as a single token, which is fed into the first layer of the LLM together with the tokens of the query. This means the context document itself does not need to be stored in external memory.

2 Results

In Section 2.1 we explore the compression, encoding, recall, consolidation, and forgetting of narrative events. We suggest how each event is efficiently encoded in the hippocampus as a conceptual code plus surprising details, from which the neocortex reconstructs the event during recall via 'retrieval-augmented generation', with these reconstructions gradually consolidated into neocortex, where interference from new consolidation produces forgetting. We compare the simulated data with the changes observed in human memory over time.

In Section 2.2 we explore the influence of semantic memory on episodic memory, and show that characteristic distortions towards the 'priors' of the network arise as a consequence, which we compare to the distortions seen in human memory. In Section 2.3 we show that consolidation of memories into the neocortical network enables learning of shared structure, which can be used to make new inferences. We simulate two non-linguistic relational inference tasks, involving the prediction of new spatial or family relationships, with the LLM trained from scratch in each case. We also show that inference is supported by the mapping of new tasks to a learned representation in the hidden layers.

Throughout these simulations, we envisage the hippocampal network as a sequential variant of the modern Hopfield network (*49, 50*), but do not simulate the mechanisms of encoding and retrieval in hippocampus in order to focus on its interactions with neocortex. For further details see Section 3.3, SI. Note that sequences are repeated many times during training, reflecting offline replay rather than online learning from a single exposure.

2.1 Modelling hippocampal-neocortical interaction as retrieval-augmented generation

We begin by simulating the compression, encoding, and consolidation of memories, and the differing mechanisms for recall over the course of a memory's lifespan.

2.1.1 The compression and encoding of episodic memory

In our model memories are compressed for more efficient storage (*51, 52*) by generating gists of events in vector form which are encoded together with unexpected details in the hippocampus. This explains not only conceptual representations in hippocampus (*53*), but also the fact that memories are schematised from the time of encoding (*54, 55*).

Retrieval-augmented generation is used to reconstruct memories from this compressed form, with hippocampal traces retrieved based on relevance to the current task or prompt, and loaded into the context of the generative network. This minimises the hippocampal storage of redundant information that is already well-predicted by the generative network (see also (*27*)). In other words, this hippocampal representation allows the neocortical network to (re)generate experiences in a more efficient way, and the conceptual representations are optimised to enable generation. We simulate this by following recent LLM research on 'context compression', aimed at mitigating the practical problems of dealing with very long prompts by transforming them into vectors which capture their meanings (*32, 56*).

Here we use the Mistral-7B-Instruct-v0.2 (*29*) LLM to represent the neocortical network, together with pre-trained weights from (*32*) to transform episodes into compressed xRAG vectors. This is one of several approaches to context compression which learn a compact vector representation of text (see also (*56*)), aligning with evidence that hippocampal traces include conceptual representations, and does not require any special training of the LLM itself - just additional weights that map the vector representation of a sequence (i.e. the output of a 'text embedding' model) to a token in the input layer of the LLM (see Figure 1f).

To illustrate how gist, detail, and semantic memory interact, we simulate the encoding and recall of narrative events (Figure 2), using 500 stories from the ROC Stories dataset (*57*). To encode a story in compressed form, its embedding is obtained from the text embedding model (*58*), from which the xRAG

vector is derived via the xRAG weights. Each story is encoded as its xRAG vector plus surprising details. To identify surprising details, all phrases are extracted from the text, then the perplexity (a measure of prediction error) of each phrase given the xRAG vector is calculated. In other words, for each phrase, the model is prompted with the xRAG vector and the phrase in question. Then the phrases are ranked in descending order of perplexity, with the top n phrases stored.

Given a query as the input to recall, the neocortex 'searches' the hippocampus for relevant traces. This is simulated by producing an xRAG vector for the query (as described above), and then by retrieving the trace(s) with the nearest xRAG vector(s). The context given to the LLM consists of the xRAG vector, the details, and a prompt describing the task. Note that the xRAG vector only occupies the space of a single token in this input. The final recalled story is the output of the LLM.

Figure 2e and Table S1 show examples of encoded and recalled stories. For instance, a story about a girl's visit to an aquarium is encoded as its gist vector and a surprising detail (that the girl took a nap on the journey). The model neocortex reconstructs an approximation of the original story from the gist vector and the detail, with a few semantic intrusions added to the story, e.g. 'Melody's favorite part was the shark tank'; these intrusions are shaped by the expectations of the model neocortex, as in (*20*). This hints at how gist, detail, and semantic memory can interact in recall.

The encoding of gist plus details was also compared to three baselines: encoding the story in full detail, encoding only the gist, and simply imagining the continuation (i.e. without using the model hippocampus at all). The mean size per memory (in tokens) and the accuracy of recall (the cosine similarity between the embeddings of the original and recalled story) were compared across these conditions, showing a trade-off between efficiency and accuracy in memory in Figure 3a. The gist plus details encoding strikes a balance between these objectives.

2.1.2 Modelling consolidation and forgetting

In our model sequential memories are initially encoded in the hippocampus in compressed form, reconstructed by the neocortex through 'retrieval-augmented generation' (*31*), and then gradually consolidated into a generative model in neocortex, through prediction error minimisation on the replayed memories. This provides a mechanism for the training of neural 'foundation models' (*24*), which can both capture specific memories *and* learn more general schematic knowledge to enable problem solving.

500 stories from the ROC Stories dataset (*57*) were encoded in gist form in the model hippocampus as described above. To simulate consolidation, the memories were reconstructed from these traces, and the generative network was trained on the result (using low-rank adaptation; (*59*)). As the story is consolidated, the neocortex's reconstruction moves closer to the encoded story in embedding space (Figure 3b), but distortions in the reconstruction are still observed. (See Section 2.2 for a more extensive exploration of distortions in our model.)

To simulate forgetting in the neocortical network, we tested the recall of the consolidated memories as the network was trained on new stories (representing the consolidation of new memories, in the absence of hippocampal reminders of the old memories). Over successive stages of new memory consolidation, the neocortex's reconstruction moves away from the encoded story in embedding space (Figure 3c). See Table 1 for an example of the same event's neocortical reconstruction accuracy improving during consolidation and declining during forgetting.

Putting these results together suggests an account for the inaccuracy of human episodic memory. To start with, the encoded version of an event is already highly compressed, relying on schemas from the neocortex. Then as consolidation progresses, memory becomes less dependent on the hippocampal trace, and more dependent on the neocortical reconstruction. We assume that the hippocampal trace is no longer maintained once consolidated into neocortex, with consolidation continuing until the trace can be *approximately* reconstructed by neocortex, rather than until it can be *perfectly* reconstructed. The consolidated version is therefore

even less accurate than the encoded version. In addition, once fully consolidated into the neocortical model forgetting occurs due to interference.

Stage	Recalled text
Original	Tina loved biology. She wanted to be an intern. She wanted to intern for a bio lab. So she applied. To her delight she got it.
Encoded (0 details)	Tina loved biology. She applied for an internship at the lab and got it. She was very excited and worked hard. She was very passionate ...
Consolidation (early)	Tina loved biology. She was very interested in it and wanted to learn more. She went to the library and checked out a book on biology.
Consolidation (late)	Tina loved biology. She applied for an internship at the lab and got it. She was very excited and worked hard. She was very passionate ...
Forgetting (early)	Tina loved biology. She applied for an internship at the lab. She was really nervous. She prepared but was not sure it was enough.
Forgetting (late)	Tina loved biology. She wanted to be a biologist. She worked very hard in school. She went to a college to get good grades. She got ...

Table 1: Text for a randomly selected story across six stages: the original story, the memory as reconstructed from the compressed hippocampal trace, the neocortical memory early and late in consolidation, and the neocortical memory early and late in forgetting.

2.2 Modelling the interplay of semantic and episodic memory in consolidation

Section 2.1 sketched out how narrative memories change over time, but now we explore the effect of existing schemas on memory in more depth. Here we investigate how the model distorts memories towards expectations, consistent with human data.

2.2.1 Gist-based distortions and prior experience

In the Bartlett experiment (*20*), students heard a story called 'The War of the Ghosts' and were asked to recall it after different time intervals. The recalled story was distorted to be more consistent with the students' background knowledge of the world, with distortion increasing over time (*61*). To simulate encoding, the story was compressed as described above (here with no poorly predicted details). To simulate consolidation, Mistral 7B was then trained on its own reconstruction of the encoded story. Recall of the story was explored by giving the network its first few words ('One night two young men from Egulac'), and inspecting the predicted continuation. The consolidated story displayed even more distortion than the encoded version (Figure 4e), suggesting that consolidation of a compressed memory is an 'approximation of an approximation'.

To explore the effect of the model's 'priors' (i.e. its expectations based on previous learning) on recall of narratives, the story from (*20*) was consolidated in addition to 'background data'. A dataset of Wikipedia content (*62*) was used, with five categories of article selected ('Politics', 'Health', 'Universe', 'Sport', and 'Nature'). The training data for each model was made up of 1000 articles sampled from the relevant category (with the first 1000 characters of each article taken) plus the Bartlett story (with the original story used to isolate the impact of consolidation). The model was trained for 10 epochs on this combined dataset to simulate consolidation, where each epoch is one full presentation of the training set. The temperature for sampling continuations from the model was also varied between 0 and 1.5 (see Section 3.3).

When the Bartlett story is 'consolidated' into the generative network memory distortions are observed, as in the human data (see Table S5). Distortions in recalled stories reflect the 'priors' of the generative network. The word clouds in Figure 4f show that new words added to the story (i.e. 'semantic intrusions') are representative of the background dataset used, and Figure 4a and b show that the recalled stories move closer towards the background dataset in text embedding space. See also Table 2 for selected examples

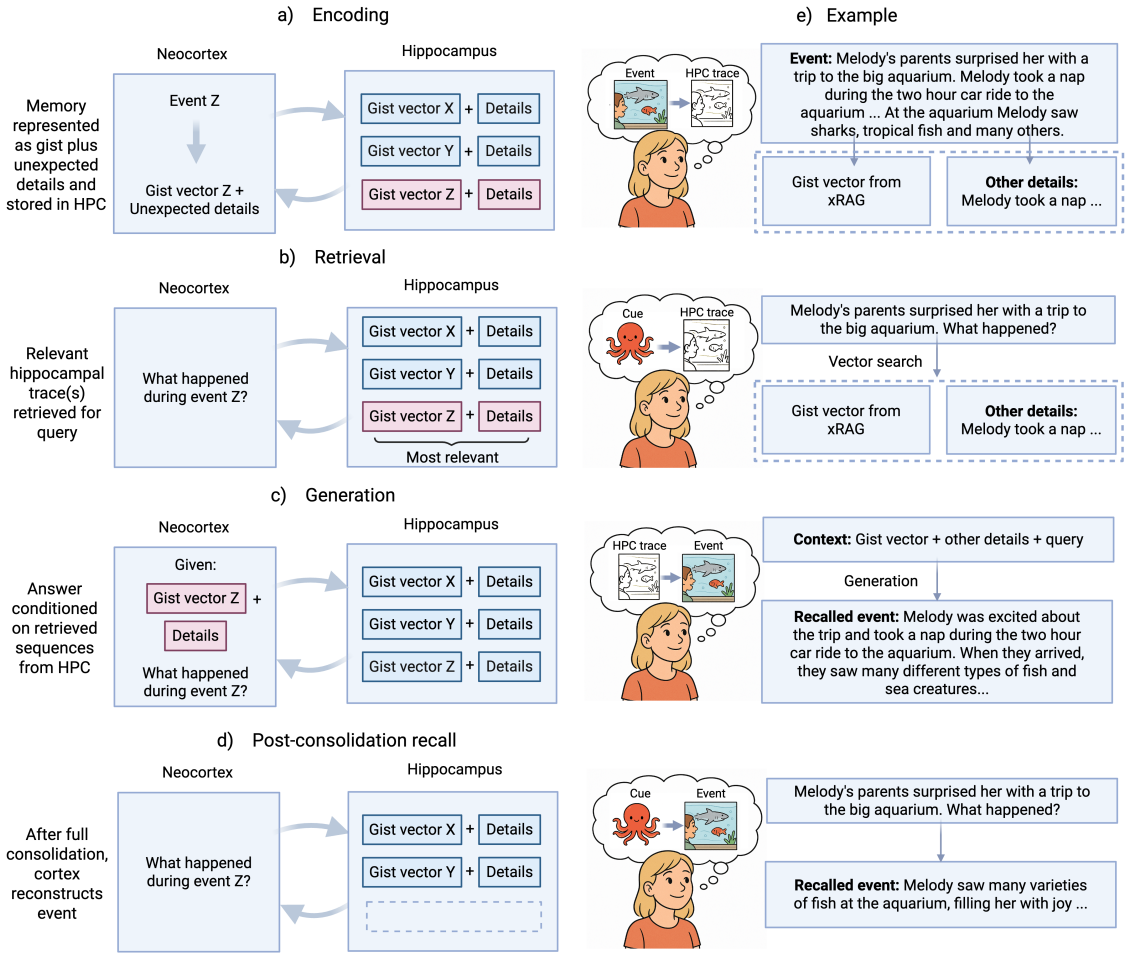

Figure 2: A model of hippocampal-neocortical interactions as retrieval-augmented generation with stories. a) The neocortex generates a gist vector for the narrative, which is encoded in the hippocampus together with unpredictable details. b) Given a query as the input to recall, the neocortex searches the hippocampus for relevant traces, e.g. by finding nearby gist vectors in this simulation. c) The generative network in neocortex produces an 'answer' conditioned on the retrieved hippocampal trace(s). d) Once a story is fully consolidated, the hippocampal trace is no longer needed to approximately reconstruct the event. e) The stages of encoding and recall are shown for an example story from the ROC Stories dataset (*57*), with Mistral-7B-Instruct-v0.2 used to simulate the neocortex, and xRAG vectors (*32*) representing the conceptual gist.

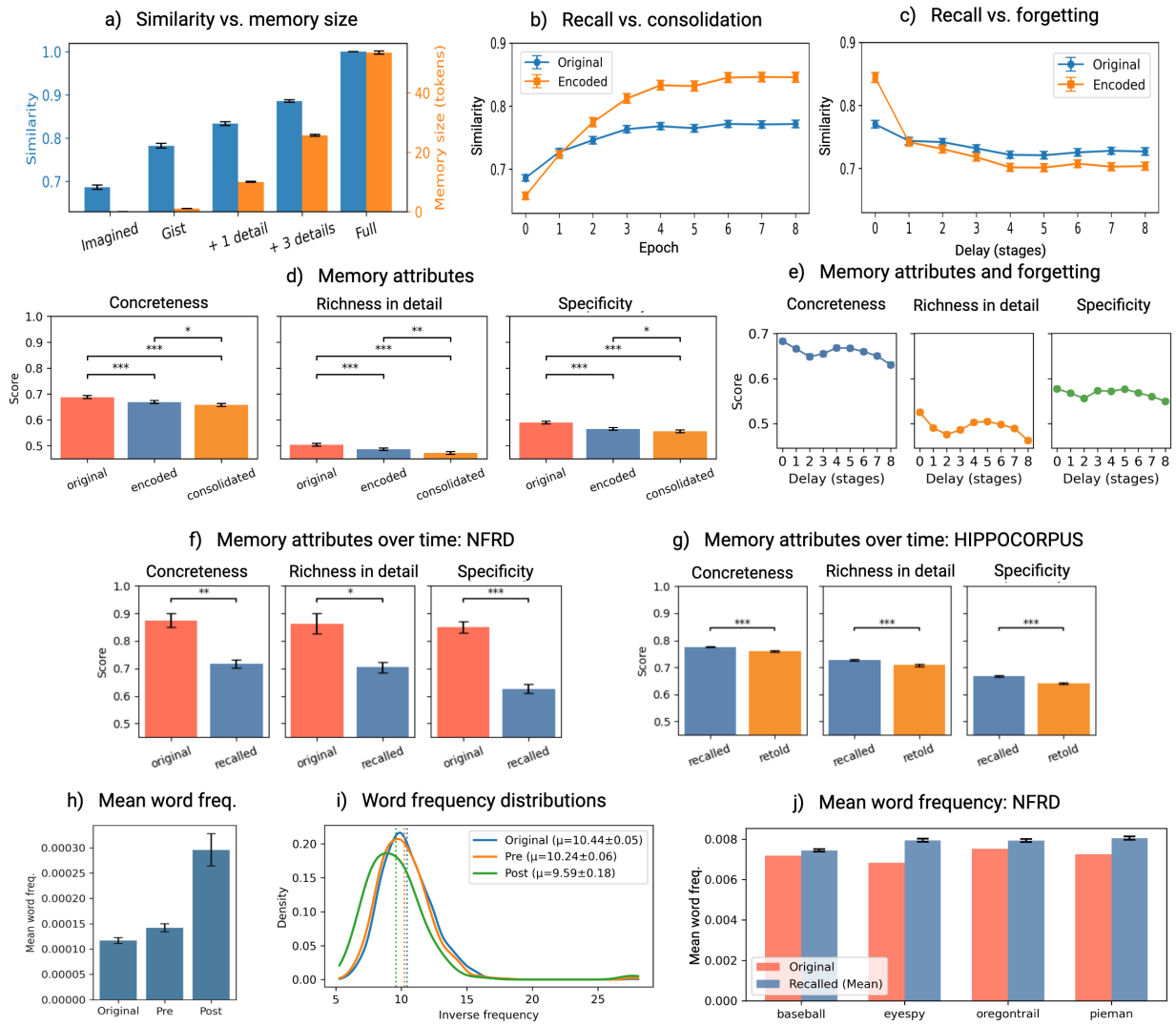

Figure 3: a) The size of the memory trace stored in the model hippocampus (in tokens) and recall accuracy (cosine similarity between original and recalled stories' embeddings with 'all-MiniLM-L6-v2'; (*60*)) for five variants. From left to right: simply imagining the event, storing only the gist vector, storing the gist vector with one surprising detail, storing the gist vector with three surprising details, and storing the event in full. b) As the story is consolidated, the neocortex's reconstruction moves closer to the encoded story in embedding space. c) As new memories are consolidated into neocortex, forgetting occurs (we assume the hippocampal trace has faded away). Over the course of new memory consolidation, the neocortex's reconstruction moves away from the encoded story in embedding space. d) Memory concreteness, richness in detail, and specificity (as scored by gpt-4o-mini) for the three subsets in our simulation: the original ROC Stories data (*57*), the memories based on the compressed hippocampal encoding, and the memories reconstructed by neocortex alone after consolidation. e) Memory concreteness, richness, and specificity decline as forgetting proceeds. f) As in d) but for original and immediately recalled stories in the Naturalistic Free Recall Dataset (NFRD; (*19*)). g) As in d) but for two subsets of the HIPPOCORPUS (*18*) dataset: 'recalled' stories, and the same stories 'retold' at a later date. h) The mean word frequency (across English language text, computed with the 'wordfreq' Python library) for words in the original, pre-consolidation, and post-consolidation stories. i) The distributions of word frequencies for these categories, showing a shift towards more common word choices after consolidation. j) As in h) but for the original and recalled versions of the four NFRD stories.

Category	Recalled story extract
Nature	It was very quiet, so quiet you could hear the grass growing.
	The waves were so high that they tossed the canoe about as if it were a toy upon the water.
	When they were out on the beach hunting seals, they saw a huge octopus.
Sport	The previous chief had been killed in a wrestling match.
	They cut themselves two canoes and provided themselves with paddles, and pushed off from the shore.
	[The] strongest warriors must make the dangerous trip down the river.
Politics	The mayor passed the message on to the trader at Egulac.
	The chief decided to avenge their death.
	[They] tried to negotiate for their release.
Health	[The] boy says that his tribe has been decimated by a disease.
	All of a sudden the water was full of blood.
	He found the bones of his companion picked clean by wolves.
Universe	The two young men set out in their canoe to hunt seals at the time of the full moon.
	They prayed to Klonos, the Klingon god of the sea, for help.
	[They] were supposed to return before dark, but they weren't back by nightfall.

Table 2: Extracts from recalled stories for different categories (temperature = 1.0), showing how semantic intrusions reflect the 'priors' of the generative network.

of semantic intrusions from the five models. Whilst distortion occurs even with 'greedy decoding', more 'semantic intrusions' are observed at higher temperatures, as Figure 4c and Table S5 show.

2.2.2 Event extension and contraction

Distortions can apply to the structure as well as the content of narratives. Event extension is the tendency to extend certain events in memory, while event contraction is the tendency to curtail them. Raykov et al. (*21*) demonstrate this with three types of video: complete videos end at a natural event boundary, incomplete videos are curtailed before this point, and updated videos are extended beyond it. They find that incomplete videos are often extended in memory (extension errors), while updated videos are often shortened (omission errors), and show that this is observed at long and short delays (Figure 4g).

We simulate the effect of consolidation in (*21*), using simple stories in text form (*63*). The majority are 'typical' stories, unmodified from the ROC Stories (*63*) dataset, whereas an 'incomplete' story has characters removed, and an 'updated' story has characters added, so it continues beyond its natural ending. The data consisted of 200 typical stories, 200 incomplete stories, and 200 updated stories, where 'updated' and 'incomplete' stories are all the same length. First the stories were encoded in compressed form as described above, and pre-consolidation recall was tested by giving the first 150 characters of each story as a query for retrieval-augmented generation. Then the *reconstructions* of these compressed memories were replayed to simulate consolidation; Mistral 7B (*29*) was fine-tuned on the reconstructed stories for three epochs, and its recall tested. Both before and after consolidation, omission errors are more common in updated than incomplete stories, while extension errors are more common in incomplete than updated ones (Figure 4h). By bringing the story to a more natural conclusion or by removing an incongruous ending, the stories are made more similar to the network's expectations. See Table 3 for examples. This mirrors findings on boundary extension and contraction in memory for images (*64, 65*); in both cases, memories are distorted towards a prior encoded in the network by its previous 'experience', consistent with Bayesian views of memory (*27,42*).

2.2.3 Comparison to changes in human narrative memory over time

The HIPPOCORPUS dataset (*18*) contains participants' written descriptions of recalled autobiographical events, with a subset of the recalled events retold months later. We analysed it to track how concrete vs. abstract, how rich vs. poor in detail, and how specific vs. general memories were for more recent and more

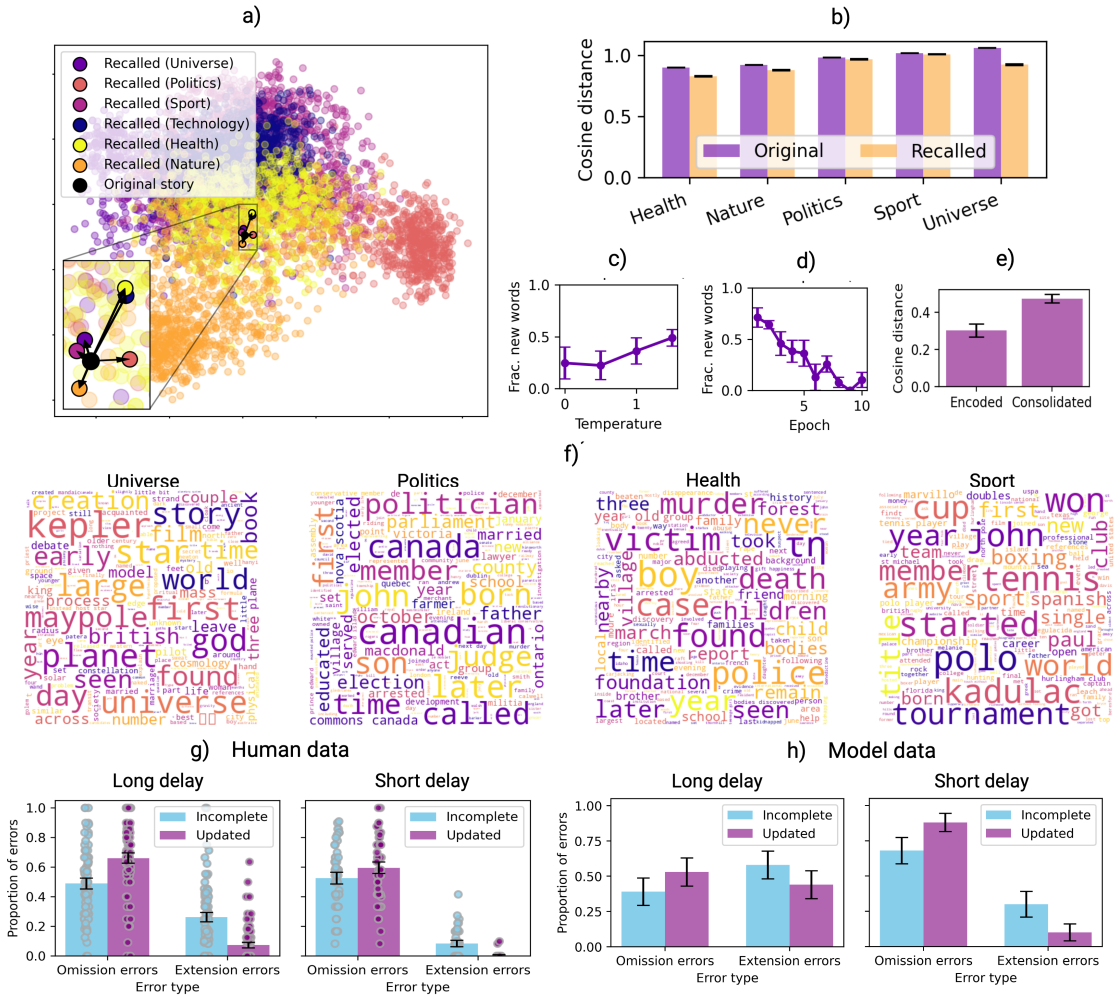

Figure 4: The effect of prior knowledge on narrative distortions. a) Mistral 7B is fine-tuned on Wikipedia content (*62*) from each of five categories plus the Bartlett (*20*) story. The embeddings of the training data plus the recalled story for each model (at a temperature of 0.5) are obtained, using 'all-MiniLM-L6-v2' (*60*). These embeddings are projected into 2D with PCA, showing that recall moves each story (see legend) towards its training data (points of the same colour). b) The cosine distance (1 - cosine similarity) between the mean embedding for each category and the original story (purple) and recalled stories (orange). Recalled stories become more similar to the background dataset. Error bars give the SEM for the 200 recalled stories sampled (but are zero for the original stories, as there is only one). c) The effect of temperature when sampling continuations on the number of new words in the recalled Bartlett stories ('semantic intrusions'). Error bars give the SEM. Results are for the model at epoch five. d) The effect of epoch on the number of new words in the recalled Bartlett stories, where an epoch is one presentation of the entire training dataset. This shows the effect of increased learning of the original story. Error bars give the SEM. Results use a temperature of 0.5. e) The mean cosine distance between the recalled and original stories before and after consolidation (at a temperature of 0.5, and with ten epochs of consolidation). f) Word clouds showing 'semantic intrusions' (produced with the 'wordcloud' Python package, with common words, i.e. 'stopwords', excluded) at a temperature of 0.5, aggregated across five trials. g) Event extension and contraction results, showing omission errors and extension errors after long (left) and short (right) retention intervals, adapted from Raykov et al. (*21*) with permission. h) The proportion of stories (200 per type) recalled with omission and extension errors, both before consolidation (cf. short delay in f) and after (cf. long delay). As in part f), error bars give the 95% confidence intervals of the mean.

Type	Original	Recalled
Typical	Stephen wanted a cardigan for his birthday. Sally bought one for him. She wrapped it up. She wrote a note. She gave it to him! The end.	Stephen wanted a cardigan for his birthday. Sally bought one for him. She wrapped it up. She wrote a note and gave it to him! The end.
Incomplete	Floyd was enjoying a summer day of inspecting cicada shells. He found many of them and hung them on his shirt like decorations. He squatted at a large tree's roots when he spotted a really nice one. When he picked it up, t The end.	Floyd was enjoying a summer day of inspecting cicada shells. He found many of them and hung them on his shirt like decorations. He squatted at a large tree's roots when he spotted a really nice cicada. He decided to add it to his collection. The end.
Updated	Kris lost her job. She was sad at first. She looked for a new job. She got an interview. Kris got a new job! Jack had always wanted his own dinosaur. He asked Santa Claus for one every year. He never would get one and beca The end.	Kris lost her job. She was sad at first. She looked for a new job. After an interview Kris got a new job! The end.

Table 3: Event extension and contraction examples before consolidation. The selected incomplete and updated stories (which are of the same length) display extension and omission errors respectively. The model is prompted with the first 150 characters of each story.

remote episodic memories. We then compared the original, encoded, and consolidated versions of stories in our simulations using the same metrics. This was achieved by using an LLM as an annotator (gpt-4o-mini via the OpenAI API).

Figure 3g shows that the more remote memories are less concrete, rich in detail, and specific, in line with many similar findings. For comparison, we simulated the process of encoding and consolidation in our model with the ROC Stories dataset (*57*) as described above, because the HIPPOCORPUS dataset does not contain 'ground truth' descriptions of the remembered events. Figure 3d shows that the encoded memories are significantly less concrete, detailed, and specific than the original stories, and the consolidated stories score even lower on these attributes. (For both the HIPPOCORPUS and simulated data, does reduction in these attributes simply reflect *less* of the memory being recalled, without any interesting changes to its content? In Tables S3 and S4, SI, we show that this is not the case by controlling for length, confirming that there are still significant differences.) In addition, Figures 3h and i show that more commonly used words occur in the remembered stories. Furthermore, memory concreteness, richness, and specificity decline as forgetting proceeds (Figure 3e).

The Naturalistic Free Recall Dataset (NFRD; (*19*)) consists of transcripts for four stories recalled by more than one hundred participants each, immediately after listening to the story. We repeated the analysis above for the original and recalled stories for comparison to the original and encoded versions in our simulation, observing a significant decrease in those attributes (Figure 3f), and furthermore the use of more common words in the recalled NFRD stories (Figure 3j). (We note that schematisation is not the only explanation for vocabulary changes; even with a perfect memory for an event, one might paraphrase its content in more familiar language.)

2.3 Neocortical learning of sequential structure for inference and planning

Next we investigate how consolidation develops generalisable knowledge. To enable simulations of training from scratch, we use a smaller LLM architecture, GPT-2 Medium (*23*), and simplify the model by storing memories in full rather than compressed form.

The ability to infer and anticipate future relations and stimuli improves with consolidation (*66–68*),

leading to downstream improvements on a diverse range of tasks even as perceptual details fade. Next we demonstrate how the neocortical network can extract sequential structure from memories through consolidation, using the example of structural inference, and illustrating the model's application to non-linguistic stimuli. A spatial example of structural inference is the finding of shortcuts, as this relies on the common structure of space, and a non-spatial example is inferring that A is the grandfather of C from the knowledge that A is the father of B, and B is the father of C, which relies on the common structure of family trees. The relations in these tasks can be seen as edges in graphs, so that multiple tasks with a common transition structure can be simulated using a fixed graph with different nodes for each task (*22*).

We simulate how consolidation enables structural inference by the neocortical network. We consider inference in two types of graph: a spatial graph and a simple family tree graph (as in (*22*)). We trained GPT-2 models on random walks on these graphs, corresponding to sequences of observations encoded in the hippocampus, and then tested the models' inference abilities on novel sequences with the same underlying structure.

In the spatial graph, a three-by-three grid represents a simple 2D environment, where the nine nodes are locations and the edges between them ('NORTH', 'EAST', 'SOUTH' and 'WEST') are possible movements (Figure 5a). Whilst each graph's structure is the same, nodes are labelled with names (random pairs of letters) to represent arbitrary features at a particular location. Trajectories through the environment are walks on the resulting directed graph, which are represented as strings such as 'ab EAST wd SOUTH ea WEST hn'. (See graph transformers for an alternative approach; (*69*).)

The family tree graph has a simple structure for illustrative purposes, consisting of two children, their parents, and two sets of grandparents. See Figure 5d. We model this as a directed graph with edges 'PARENT_OF', 'CHILD_OF', 'SPOUSE_OF', 'SIBLING_OF', 'GRANDPARENT_OF', and 'GRANDCHILD_OF'. As in the spatial case, all graphs have the same structure, but each graph has different names assigned to its nodes. Walks on the graph are represented by strings such as 'lk PARENT_OF nd SIBLING_OF re'.

In each case, we created 500,000 graphs with the same structure but randomly chosen labels (pairs of letters) for the nodes. Three random walks of 50 transitions were sampled from each graph to create the training data, representing sequences of observations that might be experienced, encoded in the hippocampus, then replayed offline. GPT-2's medium-sized architecture was then trained from scratch for one epoch (i.e. the full dataset was presented once). After training, we tested inference of the next node given the sequence so far, using greedy decoding of the output probabilities. For example, the next item after 'uq NORTH sx EAST tp SOUTH ec WEST' can be inferred to be 'uq', and the next item after 'qk PARENT_OF xm PARENT_OF vw GRANDCHILD_OF zq SPOUSE_OF' can be inferred to be 'qk'. The sequences used for testing were sequences of relationships in *new* family trees or spatial grids.

Figure 5c and f show the 'loss' (aggregated error on the training data) of the spatial and family tree models respectively. In both cases the loss gradually decreases, indicating improved ability to predict the next node on the set of graphs used for training, which corresponds to the consolidation of previous experience. Tables S6 and S7 and Figure 5g show good performance on a range of *novel* structural inference tasks, involving environments which have not been experienced before. Simpler inferences include inferring that going one step west then east takes one back to the original location, e.g. the correct continuation of 'ab EAST cd WEST' is 'ab'. Similarly, a correct continuation of 'ab PARENT_OF cd CHILD_OF' is 'ab'. But surprisingly complex inferences are also possible, e.g. that 'kh CHILD_OF oi SPOUSE_OF tv CHILD_OF fh SPOUSE_OF xr GRANDPARENT_OF gq SIBLING_OF' is followed by 'kh'. Furthermore given a single new item the trained spatial model can 'imagine' new trajectories consistent with the learned graph structure (Figure 5h), which explore further from the starting position at higher temperatures (Figure 5i). We also tested inference on larger grids than the model was trained on to show that inference cannot solely be attributed to 'rote memory' of sequences of transitions (Figure 5j).

2.3.1 Inspecting the hidden representations supporting inference

To understand how the LLM generalises to a new environment, we inspected representations in its hidden layers. Activations (vectors of length 1024) for each token of a random walk were extracted from the innermost layer (i.e. the 12th of the 24 transformer blocks) of our model, and from the same layer of the pre-trained GPT-2 medium model released by (*23*) as a baseline. Having obtained a vector for each location, we analysed the distances between these representations. We found that the nearer the points in the environment, the closer their hidden representations (Figure 6a). This is not true of the baseline model (Figure 6b). We then ran principal component analysis across representations extracted from random walks in many different environments. Figure 6d shows that our model has learned a shared structure across episodes, whereas this structure is not present in GPT-2 medium (Figure 6c). This mirrors the finding that a shared abstract representation of different spatial environments with the same structure develops in cortex through consolidation, as (*70*) show using cross-map repetition suppression effects in fMRI. Once this abstraction of the task has developed, it supports reasoning based on information in the prompt, or 'in-context learning'. The correlation between distance in the grid and distance between the representations increases with the length of a random walk on a new graph, but is high from the outset, suggesting the model can rapidly align new stimuli to its learned map (Figure 6f).

2.3.2 Inference from memory as retrieval-augmented generation

Multiple memories can be combined to support problem solving immediately after encoding, not just after consolidation. We simulate how the generative network and hippocampal network could work together to achieve this in a way resembling RAG, as in our simulation of the process of reconstructing memories from a compressed trace above. Whilst the simulations of inference in Section 2.3 involved prompting the model with the sequence of observations so far, storing all past experience in working memory is not feasible given its low capacity (*16, 17*). In general, only a small subset of past episodes are relevant, so RAG provides a way to condition predictions on only the most useful memories. We create a 'toy example' of retrieval-augmented generation using the two models in the structural inference results above. In each case, the hippocampus encodes sequences from *new* spatial or family tree graphs, corresponding to observations in a new spatial environment or of a new family's relationships. Can recent experience of multiple short paths in a new graph be combined with structural knowledge to infer a relationship that was not observed?

100 new graphs were constructed for each task, each missing a randomly chosen edge. Short walks on each graph, which did not include the missing edge, were stored in the 'hippocampus' (simply a list of strings in this example). The network's ability to infer the destination of each missing edge was tested. For example, if the 'cd PARENT_OF ef' edge was omitted from the graph, the test would be the model's continuation from 'cd PARENT_OF'. Testing involves two stages, retrieval followed by generation: first the hippocampus is queried for relevant traces, then the generative network produces an output with the retrieved sequence and the new task in the LLM's prompt (Figure 6h). The results show that this supports structural inference immediately after encoding sequences in the hippocampus, whereas relying on either the hippocampal network or generative network alone gives worse results (Figure 6i).

3 Discussion

We propose that neocortical memory takes the form of a generative network, trained to reconstruct sequential events replayed by the hippocampus. This enables the network to acquire generalisable semantic knowledge through consolidation, to reconstruct episodes efficiently from a highly compressed hippocampal trace, and eventually to capture specific events. During recall, the hippocampus retrieves episodic details which serve as

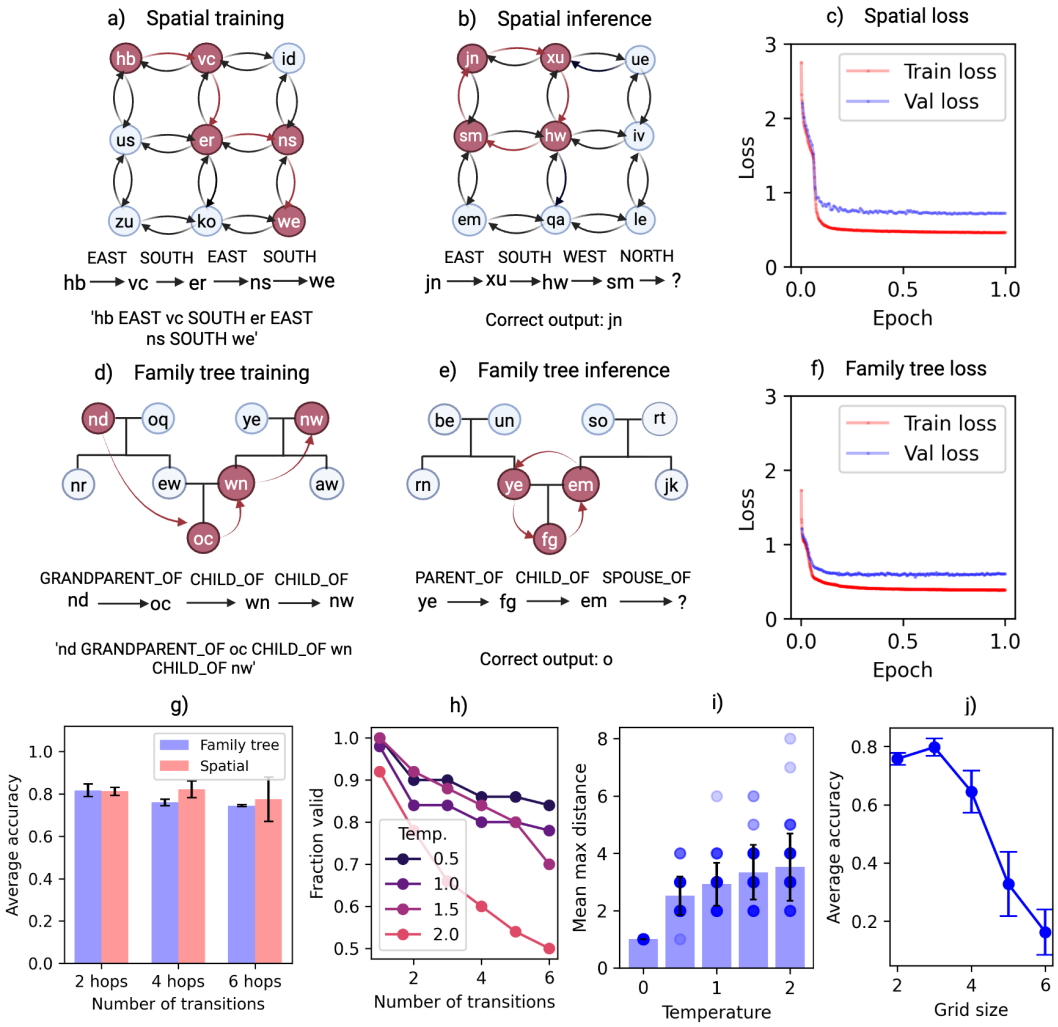

Figure 5: Learning structural regularities. a) The spatial graph task. Spatial trajectories are modelled as walks on the graph, and represented as strings. b) Testing structural inference as the ability to complete sequences on unseen graphs based on structural regularities, e.g. inferring that the next location in the sequence shown is 'jn'. c) The training and validation losses (i.e. aggregated prediction errors) of GPT-2 trained for one epoch on sequences from 500,000 spatial graphs. d) The family tree graph task. Relationships between family members are modelled as walks on the graph, and represented as strings. e) Testing structural inference, e.g. inferring that the next person in the sequence shown is 'ye'. f) The training and validation losses of GPT-2 trained for one epoch on sequences from 500,000 family tree graphs. g) Templates in which the final location can be inferred were populated with random locations, i.e. the test data are paths in novel environments not featuring in the training data. Loop completion performance for the spatial and family tree models, grouped by the number of edges ('hops') in the template. Error bars give the SEM. h) The fraction of 'imagined' spatial trajectories which are structurally valid, at a range of temperatures. A sequence is marked as invalid if different locations are found at the same inferred co-ordinates, e.g. in 'sz WEST eq EAST zr' both 'sz' and 'zr' occur at the same co-ordinates, which is contradictory. i) The maximum Manhattan distance travelled from the origin for 'imagined' trajectories of length 50 tokens (with individual data points plotted, and error bars giving the SEM). j) Generalisation to different grid sizes. The spatial inference model - trained on paths in a 3x3 grid of locations - was tested on a range of grid sizes. (The 'Grid size' axis gives the height and width of the grid in nodes.)

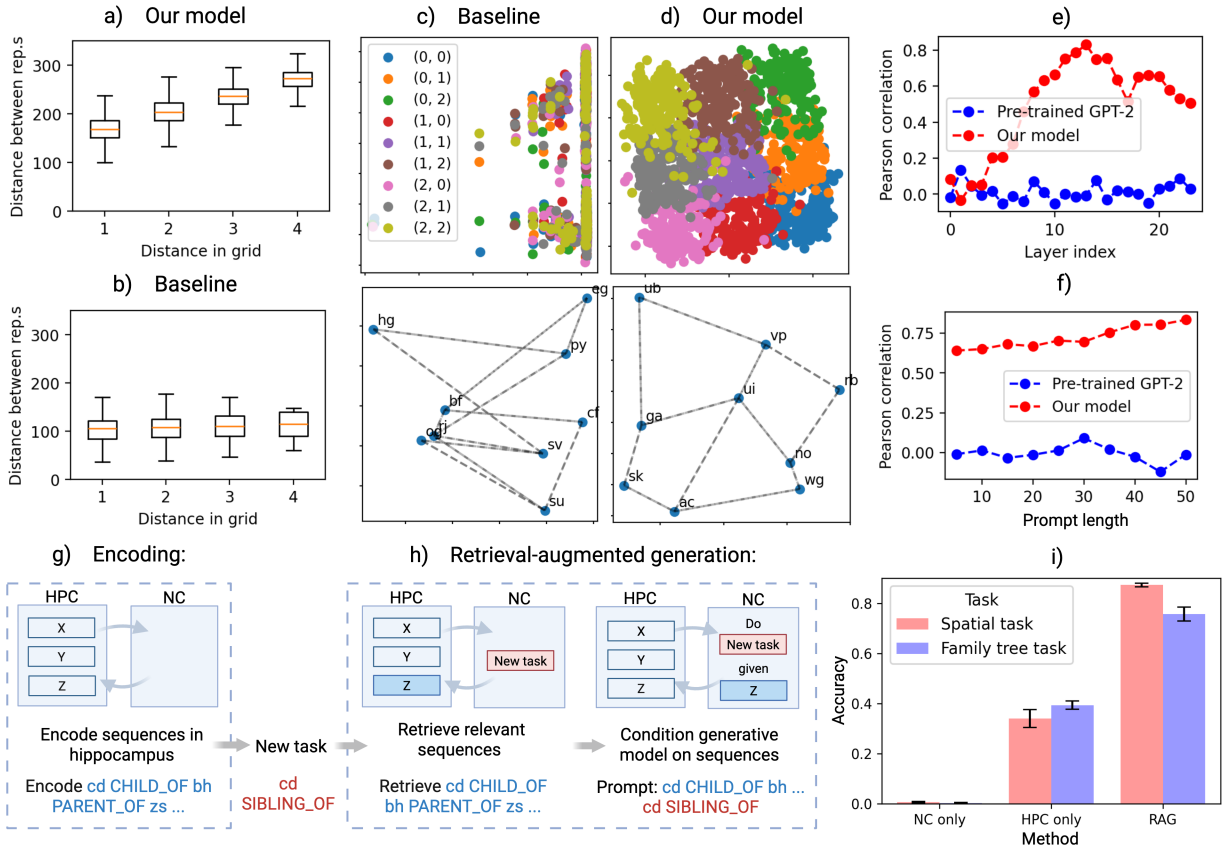

Figure 6: Inference is supported by hidden representations (a-f) and retrieval-augmented generation (RAG; g-i). a) Activations for each token of a random walk were extracted from layer 12 of our model, to obtain a vector for each location (aggregating across multiple tokens where necessary). b) As above but for pre-trained GPT-2 medium as a baseline. c-d) Principal component analysis across representations extracted from random walks in many different environments (upper row), colouring each location by its co-ordinates in the environment, so that points with the same colour are different nodes which occupy the same position in the 3x3 grid (see examples in lower row). Pre-trained GPT-2 medium is shown in c) and our model in d). e) The Pearson correlation between Manhattan distance in the environment and Euclidean distance in each layer of our model (red) and GPT-2 medium (blue). f) The Pearson correlation between Manhattan distance in the environment and Euclidean distance in L12 for differing prompt lengths, for our model (red) and pre-trained GPT-2 medium (blue). g-i) RAG applied to an inference task. g) First sequences are encoded in the hippocampus, then (before consolidation of these memories) a new task is encountered. h) Relevant sequences are retrieved from the hippocampus, and then provided as input to the generative network together with the new task. Specifically, in the spatial and family tree inference tasks the system encodes 100 walks from 100 different graphs (with a simplified structure, and each missing a single edge). Each encoded walk visits three nodes and two edges. The task is then for the system to infer the missing edge, e.g. predict the next location after 'ab NORTH' or 'cd SIBLING_OF' as shown, prior to any consolidation of the new graph. The models trained in Section 2.3 are used. i) Inference accuracy for the spatial and family tree inference tasks with RAG, and a 'HPC only' and 'NC only' baseline for comparison. The 'HPC only' baseline randomly selects one of the locations / people in the retrieved sequence. The 'NC only' baseline conditions the generative network on the task alone, without retrieving sequences from the hippocampus.

context for the neocortical network to generate predictions based on semantic knowledge. Over time, consolidation allows the neocortex to reconstruct memories independently, reducing reliance on hippocampal traces. Likewise, episodic memories guide problem-solving by providing relevant context from the hippocampus for neocortical generation.

In our model the neocortical generative network interacts with stored sequences retrieved from the hippocampus as retrieval-augmented generation, but with the 'generator' gradually trained on memories from the 'retriever' through consolidation, and the 'retriever' storing memories in a highly compressed form which can be decoded by the 'generator'. Furthermore the stored traces in the hippocampal 'retriever' can combine surprising details with gists in vector form, binding together information at different levels of abstraction in memory.

This model suggests an explanation for several features of human cognition, such as the bidirectional influences between episodic and semantic memory; semantic memory is acquired by learning to reconstruct episodic memories, but also shapes the representation of these episodes, both in terms of their initial encoding, as reflected by the compressed representations observed in hippocampus (*53*), and their gradual schematisation (*20*). To support this, we demonstrated how distortions arise in narratives due to their encoding in compressed form, then increase with consolidation and subsequent forgetting (Section 2.1). We also showed how memory reflects priors in the generative model derived from previous experience (Sections 2.2.1 and 2.2.2), and compared changes in memory over time to human data (Section 2.2.3). Our model also explains how a number of capabilities improve through consolidation, in addition to the memorisation of 'replayed' sequences. In particular, the neocortical generative network supports the inference of new relationships from limited observations by learning an implicit structure (Section 2.3), which supports prediction of future observations in sequential planning.

In recent years, neuroscience has seen a move from a modular view of many semi-independent networks learning particular tasks to a focus on the learning of multipurpose representations. Similarly, there has been a transition in machine learning from task-specific models to larger task-general ones trained through self-supervised learning, sometimes referred to as 'foundation models' (*24*). The use of a self-supervised task - simply the prediction of the next item in the sequence - to train such models is key to their scaling, as external supervision is not required. We can think of the brain as learning neural 'foundation models' too, and this paper suggests how memory consolidation could contribute to their development (see also (*27*)). This view aligns with the growing consensus that prediction error minimisation is key to biological intelligence (e.g. (*71*)), neuroimaging evidence of large task-general networks (e.g. (*72*)), and the relative lack of external supervision in learning. We do not argue that modern LLMs mirror the brain, but that some kind of neural 'foundation model' being trained to minimise prediction error on replayed sequences, and used in conjunction with specific memories, provides a good model for a wide range of data on memory, inference and planning.

3.1 Limitations and future work

This work is primarily a model of psychological, rather than neural, data. More could be done to bridge the gap between these ideas and the growing understanding of sequence representations at a neural level. For example, the model allows consideration of navigation, which is associated with distinctive cell types such as place, grid and head-direction cells, so the connection between these different levels of explanation could be explored (see (*22, 73–76*)). Furthermore, progress in 'mechanistic interpretability' (e.g. (*77*)) enables the hidden layer activations of LLMs to be compared to neuroimaging data, allowing comparison of representations at a population level as well.

Modern language models are trained on a simple next item prediction task (*23*), mirroring how prediction error minimisation is thought to drive biological learning (*71*). But even if the objective is shared, the implementations may be very different. In particular, error backpropagation - which calculates each parameter's 'responsibility' for the network's error - is thought to be implausible because of its non-local nature (*78*).

More biologically plausible approximations to error backpropagation do exist (*78*), as well as other families of network that mimic the brain more closely (*71, 79, 80*), which could be explored in future work.

Our model captures the sequential nature of experience, but in reality each moment is 'high-dimensional', more like the images in (*27*) or (*39*) than the 'tokens' used here. Video data, whilst still far simpler than reality, better reflect this: a video is made up of a succession of frames, each of which is rich in data, and recalling it requires 'filling in the gaps' of a particular frame as well as predicting successive ones. This could be modelled by reproducing the current item (as in in (*27, 39*)) as well as predicting the next, which would require changes to the associative and generative networks. Since pattern completion of both the current stimulus and the next stimulus would occur in the associative network, a combination of autoassociative and heteroassociative connectivity might be required (*81*).

We draw on a body of research connecting memory and schema formation (*8, 12, 82*). (*8*) suggest that whilst high prediction error memory elements may be preferentially encoded in hippocampus, schema congruent information may be remembered more easily by cortex. Consistent with this account, our model could predict a U-shaped relationship between surprise and recall performance, in which moderately surprising elements - those not unpredictable enough to be encoded in detail, nor familiar enough to be captured by existing schema - are especially prone to forgetting (see (*83*)). Whilst we show that memories become more schema congruent overall in our simulations, and propose a mechanism for the storage of surprising details in hippocampus, we do not model how the retention of different memory elements depends on their predictability, and how this changes over time. Future research could compare our model's predicted pattern of memorability as a function of schema congruence with these data.

Whilst experience is continuous, it is discretised into events in memory (*84, 85*), a process known as event segmentation. Event segmentation is thought to be based on prediction error (*86–88*) (but see (*89*) for an alternative account). Whilst the sequences used in our simulations are 'pre-segmented' for simplicity, this is potentially consistent with our model, as the generative network can provide an ongoing measure of 'surprise' during perception (*87, 88, 90, 91*), e.g. as quantified by perplexity (*23*). Future work could explore the connections between these ideas.

Human learning can display leaps of insight (e.g. (*92*)), with abrupt improvements thought to correspond to the reorganisation of mental representations (*93*). Future work could explore the development of an LLM's internal representations as a task is learned, comparing this to the corresponding 'behavioural' changes, to see if sudden increases in performance correspond to sudden representational changes. This could then be compared to the dynamics of learning in humans.

Future research could also explore how retrieval, encoding, and replay processes are orchestrated. We simplify matters by using pre-segmented sequences, encoding every event with a fixed level of detail, and 'hard-coding' when the hippocampus is queried. However, one could train a 'meta-controller' to co-ordinate these processes with reinforcement learning. This might learn to use different cognitive resources when it was optimal to do so, e.g. prioritising memories for replay to maximise future performance, and only encoding memories when it is worth the cognitive cost. This could build on previous studies like (*94*), which prototypes these ideas with simpler models.

Finally, catastrophic forgetting refers to the overwriting of old knowledge by new knowledge when a neural network learns multiple tasks or distributions consecutively (*95*). Whilst gradual forgetting is expected in cognition, but catastrophic forgetting is a more dramatic interference of new with old knowledge, which is not typically observed in reality. The LLM's gradient of forgetting in Section 2.1.2 is steeper than would be expected of human memory, suggesting that the brain has mechanisms for maintaining neocortical memories over the course of a lifetime that LLMs do not. Future research could investigate these mechanisms, e.g. whether generative replay in neocortex might facilitate continual learning (*96*).

3.2 Episodic and semantic memory

In our model episodic and semantic memory combine to reconstruct the past and predict the future, but how clear is this distinction? Generative networks such as large language models (LLMs) can memorise 'event-unique' specifics as well as generalities in a single network (*28*), raising questions for the common assumption that episodic and semantic memory must depend on different computational processes. If a series of narratives representing 'episodes' are 'consolidated' into an LLM, the resulting LLM could support both memory for specific episodes and for semantic 'facts', with the latter learned as a side-effect of reconstructing the former. Not only are 'beliefs' influenced by 'episodes' the network was trained on, but the 'episodes' are reshaped by the 'beliefs', as we saw in a range of memory distortion simulations. This can be the case even for the initial hippocampal encoding of episodes in the full model, as demonstrated in Section 2.1.

How do we distinguish remote memory from imagination, if both consist largely of schema-based predictions from a generative model? The generative model's prediction error for sequences from the training data tends to be lower than for novel sequences (*97*), so the generative network could provide a measure of recognition memory even if the original hippocampal trace has faded away. However this might correspond to a 'sense of familiarity' rather than definitive knowledge that a stimulus was experienced. Thus, the presence of similar real memories (*54, 55*), or the rehearsal of imagined events (*98*), can induce false memory, presumably from a 'sense of familiarity' in the generative network. This could align with hippocampal amnesics' preserved ability to make relative judgements of familiarity (*99*).

On the other hand, episodic memories arguably need to include some details from a hippocampal trace to be recognised confidently (*99, 100*). Memories may retain some limited trace in the hippocampus proper for a long time, so the retrieval of event-unique details which have not yet been consolidated could help to distinguish real memories from imagination. The neocortical network may contain many semantic facts relating to one's personal history, e.g. that one went on a holiday to a particular place as a child, and could thus construct events consistent with these facts. However, if episodic recollection requires retrieval of unique details encoded during the event, then even an accurate reconstruction would not qualify as episodic memory.

3.3 Conclusion

We have proposed a model of the bidirectional interactions between episodic and semantic memory, in which semantic memory in neocortex relies on a generative network, trained to reconstruct sequential events replayed by the hippocampus. Episodic memories are encoded in the hippocampus in compressed form, and recalled by 'decompressing' these traces with the generative network. In other words, hippocampus and neocortex work together in recall and problem-solving to provide episodic details and predictions from semantic knowledge based on these details respectively. During consolidation, these reconstructions are then replayed to update the neocortical generative network to predict the next item in each sequence. We make use of recent advances in the large language model literature to simulate this hippocampo-neocortical interaction as 'retrieval-augmented generation' (RAG), with the addition of mechanisms to *compress* episodic memories and to *consolidate* them into the generative network. This model explains changes in the characteristics of autobiographical memories over time, and accounts for gist or schema-based distortions. As well as capturing the gist of specific episodes, the neocortical network extracts statistical patterns that generalise to new situations. We also show how the hippocampal and neocortical networks could jointly contribute to problem solving through RAG, allowing memories retrieved from the hippocampus to provide the context for prediction using the 'general knowledge' of the neocortical network. We demonstrate that, by simply learning to predict the next item, the neocortical network can acquire abstracted knowledge of commonalities across experience, which can be used to support inference and planning.

Methods

Modelling hippocampal-neocortical interaction as retrieval-augmented generation

The compression and encoding of episodic memory

In Sections 2.1 and 2.2 we use Mistral-7B-Instruct-v0.2 (*29*) to explore the compression, encoding, retrieval-augmented generation, consolidation, and forgetting of narrative events. In addition to the LLM itself, pre-trained weights from (*32*) are used to transform episodes into compressed xRAG vectors. In brief, these weights are trained so that the embedding of the text reconstructed by the LLM is as close as possible to the embedding of the original. A key advantage is that the underlying idea in (*32*) is not specific to language but could in principle work for arbitrary types of sequence. Note that xRAG weights are only compatible with a particular LLM; Mistral-7B-Instruct-v0.2 is used because the xRAG weights are trained to transform embeddings into token representations of this specific model. But in principle the context compression mechanism is compatible with any LLM.

The prompt used to recall a memory with retrieval-augmented generation is as follows, formatted as suggested in the xRAG (*32*) code. Note that the 'INST' tag denotes the instruction for Mistral 7B, which is an instruction-tuned model:

> [INST] Refer to the background document and answer the question.
>
> Background: {XRAG_TOKEN}
>
> Question: {FIRST_LINE} What happened (in detail)?
>
> Other details to include: {SURPRISING_DETAILS}. [/INST] The answer is:

Here 'XRAG_TOKEN' is the single token encoding the gist produced by the xRAG projection weights, and 'SURPRISING_DETAILS' is a comma-separated list of the n highest perplexity phrases. (The line beginning 'Other details' is omitted from the prompt when no details are provided.)

Consolidation of conceptual representations

In simulations involving Mistral 7B, the model is fine-tuned in a different way because of its size. Instead of fine-tuning the entire model, parameter-efficient fine-tuning (PEFT) methods update only a small subset of parameters. One such method is low-rank adaptation (LoRA; (*59*)), which injects trainable low-rank matrices alongside frozen pre-trained weights. We implement LoRA with the 'PEFT' Python library (*101*), applying LoRA to the attention blocks alone.

The training data is formatted as follows, in the form recommended for an instruction-tuned model such as Mistral 7B:

> [INST] {FIRST_LINE} What happened (in detail)? [/INST] {REST_OF_STORY}

The prompt used to test *consolidated* memory, i.e. memory without any retrieval from the model hippocampus, is as follows:

> [INST] {FIRST_LINE} What happened (in detail)? [/INST]

Analysing the HIPPOCORPUS and NFRD datasets

The HIPPOCORPUS dataset (*18*) consists of 2779 recalled stories, 1319 retold ones, and 2756 imagined ones (not included in these analyses). The Naturalistic Free Recall Dataset (NFRD; (*19*)) consists of 116 recalls for the 'pieman' story, 116 for 'eyespy', 113 for 'pieman', and 113 for 'oregontrail'.

The following prompt was used to label the attributes of recalled and retold items in the HIPPOCORPUS dataset, and of original and recalled items in the NFRD dataset with gpt-4o-mini via the OpenAI API:

> Your task is to score text on three metrics: how concrete (vs abstract) it is, how rich in detail it is, and how specific (vs general) it is.
>
> Return ONLY a JSON dictionary with 3 keys, each a float 0-1:
>
> {'concrete_vs_abstract': 0–1, 'rich_vs_poor_details': 0–1, 'specific_vs_general': 0–1}
>
> A higher score corresponds to more concrete, richer in detail, or more specific text.

This reflects standard practices for using LLMs for data annotation (*102*).

Simulation	Section	Training method	Model	Training data
Narratives	2.1	Further training with LoRA (*59*)	Mistral-7B-Instruct-v0.2 plus xRAG weights	Reconstructed English language text from ROC Stories dataset (*57*)
Gist-based distortions	2.2.1	Further training with LoRA (*59*)	Mistral-7B-Instruct-v0.2 plus xRAG weights	Reconstructed English language text from (*20*) and Wikipedia (*62*)
Event extension and contraction	2.2.2	Further training with LoRA (*59*)	Mistral-7B-Instruct-v0.2 plus xRAG weights	Reconstructed English language text from ROC Stories dataset (*57*)
Structural inference (spatial)	2.3	From scratch	GPT-2 medium	Walks on a graph of form 'mv SOUTH sz WEST li EAST sz . . . '
Structural inference (family trees)	2.3	From scratch	GPT-2 medium	Walks on a graph of form 'yu GRANDPARENT_OF mi SIBLING_OF vb . . . '

Table 4: Summary of models trained. Here 'reconstructed' means the version of the original data reconstructed by the LLM from its compressed encoding.

Modelling the interplay of semantic and episodic memory in consolidation

Event extension and contraction are modelled with simple stories in text form, each only a few lines long (*63*). Three types of narrative are used. Some are 'typical' stories, unmodified from the (*63*) dataset, whereas an 'incomplete' story has characters removed, and an 'updated' story has characters added, so it continues beyond its natural ending. Note that this is a challenging task as some stories could be interpreted as either incomplete or updated. To disentangle the effect of story category from the effect of length, the incomplete and updated stories were all the same length. To achieve this, text from a different random story was added

to stories shorter than average, extending them to the mean length, while longer than average stories were shortened to the mean length.

After encoding and reconstructing the stories, the Mistral 7B model (*29*) was fine-tuned on the reconstructions for three epochs and then recall was tested by giving the model the first 150 characters and observing the output.

Neocortical learning of sequential structure for inference and planning

In Section 2.3 we train sequence models (GPT-2; (*23*)) on non-linguistic sequences, choosing a smaller model to run more extensive simulations of learning structural inference from scratch. The 'medium' sized GPT-2 model is used, which has 345 million parameters including 24 transformer blocks, and is trained from scratch with randomly initialised weights. See Table 4 for a summary of the models trained and Table 5 for a summary of the hyperparameters.

Simulation	**Training method**	**Epochs**	**Learning rate**	**Batch size**	**Training data**	**Trials**
Structural inference (spatial)	Training GPT-2 medium architecture from scratch	1	5e-05	1	Three random walks of length 50 on each of 500,000 graphs	1
Structural inference (family trees)	Training GPT-2 medium architecture from scratch	1	5e-05	1	Three random walks of length 50 on each of 500,000 graphs	1

Table 5: Summary of training hyperparameters and data volumes for GPT-2 training.

We simulated how consolidation enables structural inference by the neocortical network, considering the example of a spatial graph and a simple family tree graph (as in (*22*)). We trained GPT-2 medium from scratch on random walks on these graphs, corresponding to sequences of observations encoded in the hippocampus, and then tested the models' inference abilities on novel sequences with the same underlying structure.

Note that on some of the family tree inference problems there are multiple possible answers which are consistent with, but cannot be inferred from, the sequence so far (such as the imagined family member 'ef' in 'ab CHILD_OF cd PARENT_OF ef', whereas 'ab' is the expected answer), and these are counted as incorrect, making this quite a harsh performance metric.

We simulate episodic contributions to problem solving as retrieval-augmented generation, using the two models in the structural inference results above (one trained on spatial graphs, and one on family tree graphs, such that each has learned the structural regularities of the stimuli). In each case, the hippocampus encodes sequences from *new* spatial or family tree graphs, corresponding to observations in a new spatial environment or of a new family's relationships. 100 new graphs were constructed for each task, each missing a randomly chosen edge. Short walks on each graph, which did not include the missing edge, were stored in the 'hippocampus' (simply a list of strings in this example). For each missing edge, a query of the form 'ab EAST' or 'cd PARENT_OF' was constructed for the spatial and family tree graphs respectively. In other words, if the 'cd PARENT_OF ef' edge was omitted from the graph, the test would be the model's continuation from 'cd PARENT_OF'. Testing involved two stages, retrieval followed by generation: first the hippocampus is queried for relevant traces, simply by finding sequences containing the node in the query. Then the generative network produces an output conditioned on the retrieved sequences concatenated with the sequence for the task (Figure 6h), i.e. with the retrieved sequence and the current task in the LLM's

prompt. Greedy decoding was used to generate predictions.

References

1. W. B. Scoville, B. Milner, Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry* **20** (1), 11 (1957).

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

3. J. L. McClelland, B. L. McNaughton, R. C. O'Reilly, Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review* **102** (3), 419 (1995).

4. P. Alvarez, L. R. Squire, Memory consolidation and the medial temporal lobe: a simple network model. *Proceedings of the national academy of sciences* **91** (15), 7041–7045 (1994).

5. J. Robin, M. Moscovitch, Details, gist and schema: hippocampal–neocortical interactions underlying recent and remote episodic and spatial memory. *Current opinion in behavioral sciences* **17**, 114–123 (2017).

6. V. Samborska, J. L. Butler, M. E. Walton, T. E. Behrens, T. Akam, Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nature Neuroscience* **25** (10), 1314–1326 (2022).

7. M. Moscovitch, L. Nadel, Consolidation and the hippocampal complex revisited: in defense of the multiple-trace model. *Current opinion in neurobiology* **8** (2), 297–300 (1998).

8. M. T. Van Kesteren, D. J. Ruiter, G. Fernández, R. N. Henson, How schema and novelty augment memory formation. *Trends in neurosciences* **35** (4), 211–219 (2012).

9. J. L. McClelland, B. L. McNaughton, A. K. Lampinen, Integration of new information in memory: new insights from a complementary learning systems perspective. *Philosophical Transactions of the Royal Society B* **375** (1799), 20190637 (2020).

10. M. A. L. Ralph, E. Jefferies, K. Patterson, T. T. Rogers, The neural and computational bases of semantic cognition. *Nature reviews neuroscience* **18** (1), 42–55 (2017).

11. O. Bein, C. Gasser, T. Amer, A. Maril, L. Davachi, Predictions transform memories: How expected versus unexpected events are integrated or separated in memory. *Neuroscience & Biobehavioral Reviews* **153**, 105368 (2023).

12. M. J. Sekeres, J. Schomaker, L. Nadel, D. Tse, To update or to create? The influence of novelty and prior knowledge on memory networks. *Philosophical Transactions B* **379** (1906), 20230238 (2024).

13. C. V. Dong, Q. Lu, K. A. Norman, S. Michelmann, Towards large language models with human-like episodic memory. *Trends in Cognitive Sciences* (2025).

14. S. J. Gershman, I. Fiete, K. Irie, Key-value memory in the brain. *Neuron* **113** (11), 1694–1707 (2025).

15. A. Baddeley, Working memory. *Science* **255** (5044), 556–559 (1992).

16. A. Baddeley, The episodic buffer: a new component of working memory? *Trends in cognitive sciences* **4** (11), 417–423 (2000).

17. N. Cowan, The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences* **24** (1), 87–114 (2001).

18. M. Sap, E. Horvitz, Y. Choi, N. A. Smith, J. Pennebaker, Recollection versus imagination: Exploring human memory and cognition via neural language models, in *Proceedings of the 58th annual meeting of the association for computational linguistics* (2020), pp. 1970–1978.

19. O. Raccah, P. Chen, T. M. Gureckis, D. Poeppel, V. A. Vo, The "Naturalistic Free Recall" dataset: four stories, hundreds of participants, and high-fidelity transcriptions. *Scientific Data* **11** (1), 1317 (2024).

20. F. C. Bartlett, *Remembering: A study in experimental and social psychology* (Cambridge university press) (1932).

21. P. P. Raykov, D. Varga, C. M. Bird, False memories for ending of events. *Journal of Experimental Psychology: General* (2023).

22. J. C. Whittington, *et al.*, The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell* **183** (5), 1249–1263 (2020).

23. A. Radford, *et al.*, Language models are unsupervised multitask learners. *OpenAI blog* **1** (8), 9 (2019).

24. R. Bommasani, *et al.*, On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

25. T. Brown, *et al.*, Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).

26. J. L. McClelland, T. T. Rogers, The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience* **4** (4), 310–322 (2003).

27. E. Spens, N. Burgess, A generative model of memory construction and consolidation. *Nature Human Behaviour* pp. 1–18 (2024).

28. N. Carlini, *et al.*, Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646* (2022).

29. A. Q. Jiang, *et al.*, Mistral 7B (2023), `https://arxiv.org/abs/2310.06825`.

30. W. Sun, M. Advani, N. Spruston, A. Saxe, J. E. Fitzgerald, Organizing memories for generalization in complementary learning systems. *Nature neuroscience* **26** (8), 1438–1448 (2023).

31. P. Lewis, *et al.*, Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems* **33**, 9459–9474 (2020).

32. X. Cheng, *et al.*, xrag: Extreme context compression for retrieval-augmented generation with one token. *arXiv preprint arXiv:2405.13792* (2024).

33. D. Li, *et al.*, Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110* (2022).

34. R. C. Atkinson, R. M. Shiffrin, Human memory: A proposed system and its control processes. *The psychology of learning and motivation* **2** (1968).

35. D. Marr, A theory for cerebral neocortex. *Proceedings of the Royal society of London. Series B. Biological sciences* **176** (1043), 161–234 (1970).

36. D. L. Schacter, D. R. Addis, R. L. Buckner, Remembering the past to imagine the future: the prospective brain. *Nature reviews neuroscience* **8** (9), 657–661 (2007).

37. G. Buzsáki, The hippocampo-neocortical dialogue. *Cerebral cortex* **6** (2), 81–92 (1996).

38. A. Baddeley, The concept of episodic memory. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **356** (1413), 1345–1350 (2001).

39. Z. Fayyaz, *et al.*, A model of semantic completion in generative episodic memory. *Neural Computation* **34** (9), 1841–1870 (2022).

40. S. Káli, P. Dayan, Hippocampally-dependent consolidation in a hierarchical model of neocortex. *Advances in Neural Information Processing Systems* **13** (2000).

41. S. Káli, P. Dayan, Replay, repair and consolidation. *Advances in Neural Information Processing Systems* **15** (2002).

42. P. Hemmer, M. Steyvers, A Bayesian account of reconstructive memory. *Topics in Cognitive Science* **1** (1), 189–202 (2009).

43. D. G. Nagy, B. Török, G. Orbán, Optimal forgetting: Semantic compression of episodic memories. *PLoS Computational Biology* **16** (10), e1008367 (2020).

44. D. G. Nagy, G. Orbán, C. M. Wu, Adaptive compression as a unifying framework for episodic and semantic memory. *Nature Reviews Psychology* pp. 1–15 (2025).

45. B. B. Doll, K. D. Duncan, D. A. Simon, D. Shohamy, N. D. Daw, Model-based choices involve prospective neural activity. *Nature neuroscience* **18** (5), 767–772 (2015).

46. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69** (6), 1204–1215 (2011).

47. O. Vikbladh, N. Burgess, E. Russek, Systems Consolidation of Sequential Dynamics in Model-Based Planning. *bioRxiv* (2024).

48. L. Chen, *et al.*, Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* **34**, 15084–15097 (2021).

49. H. Ramsauer, *et al.*, Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217* (2020).

50. B. Millidge, T. Salvatori, Y. Song, T. Lukasiewicz, R. Bogacz, Universal Hopfield networks: A general framework for single-shot associative memory models, in *International Conference on Machine Learning* (PMLR) (2022), pp. 15561–15583.

51. H. B. Barlow, *et al.*, Possible principles underlying the transformation of sensory messages. *Sensory communication* **1** (01), 217–233 (1961).

52. H. B. Barlow, Unsupervised learning. *Neural computation* **1** (3), 295–311 (1989).

53. R. Q. Quiroga, Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience* **13** (8), 587–597 (2012).

54. J. Deese, On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of experimental psychology* **58** (1), 17 (1959).

55. H. L. Roediger, K. B. McDermott, Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition* **21** (4), 803 (1995).

56. T. Ge, *et al.*, In-context autoencoder for context compression in a large language model. *arXiv preprint arXiv:2307.06945* (2023).

57. N. Mostafazadeh, *et al.*, A corpus and cloze evaluation for deeper understanding of commonsense stories, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2016), pp. 839–849.

58. R. Meng, *et al.*, SFR-Embedding-Mistral:Enhance Text Retrieval with Transfer Learning, Salesforce AI Research Blog (2024), `https://www.salesforce.com/blog/sfr-embedding/`.

59. E. J. Hu, *et al.*, Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

60. N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics) (2019), `https://arxiv.org/abs/1908.10084`.

61. E. T. Bergman, H. L. Roediger, Can Bartlett's repeated reproduction experiments be replicated? *Memory & cognition* **27** (6), 937–947 (1999).

62. T. Ziadé, Wikipedia Topics, `https://huggingface.co/datasets/tarekziade/wikipedia-topics` (2024), accessed: 2024-05-31.

63. H. Rashkin, A. Bosselut, M. Sap, K. Knight, Y. Choi, Modeling naive psychology of characters in simple commonsense stories. *arXiv preprint arXiv:1805.06533* (2018).

64. H. Intraub, M. Richardson, Wide-angle memories of close-up scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **15** (2), 179 (1989).

65. W. A. Bainbridge, C. I. Baker, Boundaries extend and contract in scene memory depending on image properties. *Current Biology* **30** (3), 537–543 (2020).

66. J. M. Ellenbogen, P. T. Hu, J. D. Payne, D. Titone, M. P. Walker, Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences* **104** (18), 7723–7728 (2007).

67. D. Kumaran, What representations and computations underpin the contribution of the hippocampus to generalization and inference? *Frontiers in Human Neuroscience* **6**, 157 (2012).

68. H. Tarder-Stoll, C. Baldassano, M. Aly, Consolidation enhances sequential multistep anticipation but diminishes access to perceptual features. *Psychological Science* p. 09567976241256617 (2024).

69. V. P. Dwivedi, X. Bresson, A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699* (2020).

70. A. Baram, *et al.*, An abstract relational map emerges in the human medial prefrontal cortex with consolidation. *bioRxiv* pp. 2024–10 (2024).

71. K. Friston, The free-energy principle: a unified brain theory? *Nature reviews neuroscience* **11** (2), 127–138 (2010).

72. M. E. Raichle, *et al.*, A default mode of brain function. *Proceedings of the national academy of sciences* **98** (2), 676–682 (2001).

73. K. L. Stachenfeld, M. M. Botvinick, S. J. Gershman, The hippocampus as a predictive map. *Nature neuroscience* **20** (11), 1643–1653 (2017).

74. A. Banino, *et al.*, Vector-based navigation using grid-like representations in artificial agents. *Nature* **557** (7705), 429–433 (2018).

75. F. Schönfeld, L. Wiskott, Modeling place field activity with hierarchical slow feature analysis. *Frontiers in computational neuroscience* **9**, 51 (2015).

76. B. Sorscher, G. C. Mel, S. A. Ocko, L. M. Giocomo, S. Ganguli, A unified theory for the computational and mechanistic origins of grid cells. *Neuron* **111** (1), 121–137 (2023).

77. T. Bricken, *et al.*, Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread* (2023), https://transformer-circuits.pub/2023/monosemantic-features/index.html.

78. J. C. Whittington, R. Bogacz, Theories of error back-propagation in the brain. *Trends in cognitive sciences* **23** (3), 235–250 (2019).

79. P. Dayan, G. E. Hinton, R. M. Neal, R. S. Zemel, The Helmholtz machine. *Neural computation* **7** (5), 889–904 (1995).

80. R. P. Rao, D. H. Ballard, Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* **2** (1), 79–87 (1999).

81. H. Sompolinsky, I. Kanter, Temporal association in asymmetric neural networks. *Physical review letters* **57** (22), 2861 (1986).

82. B. Milivojevic, A. Vicente-Grabovetsky, C. F. Doeller, Insight reconfigures hippocampal-prefrontal memories. *Current Biology* **25** (7), 821–830 (2015).

83. J. A. Quent, A. Greve, R. N. Henson, Shape of U: The nonmonotonic relationship between object–location memory and expectedness. *Psychological Science* **33** (12), 2084–2097 (2022).

84. D. Newtson, Attribution and the unit of perception of ongoing behavior. *Journal of personality and social psychology* **28** (1), 28 (1973).

85. D. Newtson, G. Engquist, The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology* **12** (5), 436–450 (1976).

86. J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, J. R. Reynolds, Event perception: a mind-brain perspective. *Psychological bulletin* **133** (2), 273 (2007).

87. N. T. Franklin, K. A. Norman, C. Ranganath, J. M. Zacks, S. J. Gershman, Structured Event Memory: A neuro-symbolic model of event cognition. *Psychological Review* **127** (3), 327 (2020).

88. M. Kumar, *et al.*, Bayesian surprise predicts human event segmentation in story listening. *Cognitive science* **47** (10), e13343 (2023).

89. A. C. Schapiro, T. T. Rogers, N. I. Cordova, N. B. Turk-Browne, M. M. Botvinick, Neural representations of events arise from temporal community structure. *Nature neuroscience* **16** (4), 486–492 (2013).

90. S. Michelmann, M. Kumar, K. A. Norman, M. Toneva, Large language models can segment narrative events similarly to humans. *Behavior Research Methods* **57** (1), 1–13 (2025).

91. Z. Fountas, *et al.*, Human-like episodic memory for infinite context llms. *arXiv preprint arXiv:2407.09450* (2024).

92. J. Kounios, M. Beeman, The cognitive neuroscience of insight. *Annual review of psychology* **65** (1), 71–93 (2014).

93. S. Ohlsson, Information-processing explanations of insight and related phenomena. *Advances in the psychology of thinking* **1**, 1–44 (1992).

94. Q. Lu, U. Hasson, K. A. Norman, A neural network model of when to retrieve and encode episodic memories. *elife* **11**, e74445 (2022).

95. R. Hadsell, D. Rao, A. A. Rusu, R. Pascanu, Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences* **24** (12), 1028–1040 (2020).

96. G. M. Van de Ven, A. S. Tolias, Generative replay with feedback connections as a general strategy for continual learning. *arXiv preprint arXiv:1809.10635* (2018).

97. N. Carlini, *et al.*, Membership inference attacks from first principles, in *2022 IEEE Symposium on Security and Privacy (SP)* (IEEE) (2022), pp. 1897–1914.

98. E. F. Loftus, J. E. Pickrell, The formation of false memories (1995).

99. E. Migo, D. Montaldi, K. A. Norman, J. Quamme, A. Mayes, The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Quarterly Journal of Experimental Psychology* **62** (6), 1198–1215 (2009).

100. A. P. Yonelinas, L. J. Otten, K. N. Shaw, M. D. Rugg, Separating the brain regions involved in recollection and familiarity in recognition memory. *Journal of Neuroscience* **25** (11), 3002–3008 (2005).

101. S. Mangrulkar, *et al.*, PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods, `https://github.com/huggingface/peft` (2022).

102. Z. Tan, *et al.*, Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446* (2024).

103. T. Wolf, *et al.*, HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

104. J. Alammar, The Illustrated Transformer (2018), `https://jalammar.github.io/illustrated-transformer/`, accessed on March 19, 2024.

105. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

106. J. Alammar, The illustrated GPT-2 (2019), `https://jalammar.github.io/illustrated-gpt2/`, accessed on March 19, 2024.

107. D. Krotov, J. Hopfield, Large associative memory problem in neurobiology and machine learning. *arXiv preprint arXiv:2008.06996* (2020).

108. D. Krotov, J. J. Hopfield, Dense associative memory for pattern recognition. *Advances in neural information processing systems* **29** (2016).

109. H. T. Chaudhry, J. A. Zavatone-Veth, D. Krotov, C. Pehlevan, Long sequence Hopfield memory. *arXiv preprint arXiv:2306.04532* (2023).

110. M. Tang, H. Barron, R. Bogacz, Sequential Memory with Temporal Predictive Coding. *arXiv preprint arXiv:2305.11982* (2023).

111. T. Salvatori, *et al.*, Associative memories via predictive coding. *Advances in Neural Information Processing Systems* **34**, 3874–3886 (2021).

112. N. Burgess, G. J. Hitch, Memory for serial order: A network model of the phonological loop and its timing. *Psychological review* **106** (3), 551 (1999).

113. M. W. Howard, M. J. Kahana, A distributed representation of temporal context. *Journal of mathematical psychology* **46** (3), 269–299 (2002).

Acknowledgments

We thank Oliver Vikbladh, Misun Kim, Steve Fleming, Kris Jensen, and Mathias Sablé-Meyer for their insightful comments on the manuscript.

Funding: Funding support for this work was received from a Wellcome Principal Research Fellowship 'Neural mechanisms of memory and prediction: Finding structure in experience' (222457/Z/21/Z) (N.B.) and a Wellcome Collaborative Award 'Organising knowledge for flexible behaviour in the prefrontal-hippocampal circuitry' (214314/Z/18/Z) (N.B.).

Author contributions: Conceptualization: ES, NB. Methodology: ES, NB, Investigation: ES. Visualization: ES. Supervision: NB. Writing—original draft: ES, NB. Writing—review & editing: NB.

Competing interests: There are no competing interests to declare.

Data and materials availability: Code for all simulations can be found at https://github.com/ellie-as/hippocampal-neocortical-RAG. All datasets used are publicly available.

Supplementary information

Further model details

Modelling sequence learning

The simulations described above use autoregressive sequence models to represent the generative networks trained through hippocampal replay. Specifically transformer-based deep neural networks for text generation are used.

The primary goal during training is to adjust the model's parameters through maximum likelihood estimation, so that the probability it predicts for the true next item in each sequence, based on the items so far, is as high as possible. In other words, the network's weights are updated to predict the probability distribution of the next item as accurately as possible, a process known as causal language modelling. The training data for the original GPT-2 model is the WebText dataset of online content. Once the model is trained, it predicts the probability distribution across all items given the items so far. To generate a sequence one can sample from this distribution or simply take the most probable item at each step. The equation below gives the probability of a sequence x as a product of conditional probabilities of its items:

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, ..., s_{n-1})$$

We use the Transformers Python library (*103*) in all simulations. The first stage of using an LLM is to prepare the inputs with tokenisation. A tokeniser maps commonly occurring chunks of characters to IDs (in order to look up the right token embedding in a learned embedding matrix); in the case of language tokens are often words or parts of words. The concept of tokenisation is applicable to arbitrary sequences, but for simplicity and consistency across the simulations all stimuli are converted to strings of characters (if they are not already text-based). For both Mistral 7B and GPT-2 training, the same tokeniser is used as in the pre-trained model.

What exactly does causal language modelling with a custom dataset involve? First the training data is split into blocks, and then for every block the cross-entropy loss is aggregated across all the next token

prediction tasks within the block. For example, for GPT-2 the block size (which is also the context size of the trained model) is 1024 tokens. This means that the model is trained to consider up to 1024 tokens of context when predicting the next token in a sequence. So for each block the model tries to predict the second token based on the first token, then the third token based on the first two, and so on, until it predicts the final token based on the preceding 1023.

The loss measures the difference between two probability distributions: the distribution predicted by the model and the actual distribution in the data. For each token prediction task, the actual distribution is a 'one-hot' vector with a one for the real next token and zeros elsewhere. Specifically, the cross-entropy loss for a single prediction task is calculated as the negative log probability assigned by the model to the actual next token. For a block of tokens, the total loss is the sum of the cross-entropy losses for each token prediction task within the block, and the weights of the model are updated based on this total loss. This procedure is the same whether the model is fine-tuned or trained from scratch.

The learning rate decreases over the course of training in all simulations, as per the default behaviour of the Trainer class in the Transformers Python library (*103*).

Sampling options

There are many ways to generate sequences from a trained sequence model like Mistral 7B or GPT-2. The simplest is greedy decoding, where the model selects the token with the highest probability as the next token in the sequence. However this can lead to repetitive and predictable sequences. Sampling from the learned probability distribution with a given temperature introduces randomness into the selection of the next item, and provides a way to control the model's 'imaginativeness'. Specifically, the temperature parameter determines the 'sharpness' of the distribution from which output tokens are selected, so that sequences at a higher temperature are more 'imaginative', but more likely to be nonsensical.

The equation below describes the 'softmax with temperature' function that is applied to the vector of scores for each token to produce a vector of *probabilities*. A large temperature T flattens the distribution, whereas T close to zero approximates a 'one-hot' vector, with a probability of one for the most likely token:

$$\sigma(s_i) = \frac{e^{\frac{s_i}{T}}}{\sum_{j=1}^{n} e^{\frac{s_j}{T}}}$$

Unless stated otherwise, sequences are sampled from the probability distribution of tokens with a given temperature. Beam search is a search strategy used in some simulations, which considers multiple potential paths through the model's probability 'landscape'. At each step, it keeps a fixed number (the beam width) of the most probable sequences generated so far and extends them, eventually choosing the sequence with the highest overall probability.

Further details on attention

In short, the attention mechanism works as follows (explained with reference to GPT-2): a query, a key, and a value vector are produced for each element in the sequence by learned weights. For a given element, the model computes a set of attention scores which determine how much attention should be paid to each element of the input sequence when representing this element. This is achieved by taking the dot products of the given element's query vector with the key vectors for each element. The final output is the sum of the value vectors for each element weighted by the corresponding attention scores (with some normalisation), allowing the model to aggregate information from the most relevant parts of the input. This describes a single 'attention head', but GPT-2's architecture features multiple attention blocks, each with multiple attention heads, which tend to develop different 'specialisms' through training. To be more precise the equation is given below:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

In this equation, Q, K and V represent the query, key, and value matrices, respectively. These are derived from the input data. d_k represents the dimensionality of the keys (and queries), and is used to scale the dot products in a way that leads to more stable gradients. The softmax function is applied to ensure the weights sum up to 1. When the resulting term is matrix multiplied by V, the result is the relevance-weighted sum of the vector representations in V, for each token in V. In other words, the attention mechanism enriches a token's representations with a relevance-weighted sum of the other tokens' representations. See (*104*) for a detailed description of this process.

Self-attention is a special case of attention in which elements of a sequence 'attend to' the sequence they are a part of. Specifically, 'masked' self-attention is used, meaning that the representation of a token in an attention block only 'attends to' preceding tokens. Previous models like BERT (*105*) perform much less well at text generation partly because their attention is not masked in this way.

Transformer blocks come in different varieties. GPT-2 (*23*) is a decoder-only model consisting of a stack of transformer decoder blocks. Each such block consists of a masked self-attention layer followed by a feedforward layer. In a nutshell, inference with GPT-2 works as follows: First an embedding for each token is obtained from a learned embedding matrix. A key point is that position embeddings representing the tokens' positions in the sequence are added to the token embeddings (these are required because self-attention is permutation-invariant, so otherwise information about the order of items in the sequence would be lost). Then the sequence of embeddings passes through a series of transformer decoder blocks, each further enriching the token representations so that they come to capture the meaning of the text. See (*106*) for helpful illustrations of the stages of processing in GPT-2. GPT-2 (*23*) comes in several sizes, with the number of transformer blocks depending on the size of the model. The small variant has 12 transformer blocks, the medium variant has 24 transformer blocks, and the large variant has 36 transformer blocks.

Supplementary results

Supplementary tables

Table S1 shows examples of encoded and recalled stories from Section 2.1.

Tables S2, S3, and S4 show that the patterns in concreteness, specificity, and richness in detail hold even when controlling for length. This is important because for both the HIPPOCORPUS and simulated data, there are significant Pearson correlations between length and each attribute (Table S2). For the HIPPOCORPUS data, regressing out the length, and then comparing the residual score for each attribute (i.e. the component which cannot be explained by the length) using a t-test, gives the results in Table S3. Doing the same for the simulated data gives the results in Table S4.

Table S5 shows the Bartlett story recalled at different temperatures, illustrating the approximate memorisation of the narrative in Section 2.2.1. Tables S6 and S7 give the inference accuracy for each template tested in Section 2.3.

Modelling sequence memory in the hippocampus

The mechanisms by which sequences are stored as traces in the initial hippocampal network are not the focus of this paper, and the hippocampus is not simulated explicitly in the results above. However, we now outline how the hippocampus might store sequences in the sequential model, in a way that supports replay and recall prior to consolidation.

In the static model in (*27*), a modern continuous Hopfield network (MCHN; (*49*)) represents the hippocampus, interpreted such that the feature units activated by an event are bound together by a memory

Original	Gist only	Stored detail	Gist + detail
Sarah has a bad habit. All Sarah eats is candy and sweets. She starts to notice a sharp pain in her teeth. The dentist tells her she has 6 cavities! Sarah decides she needs to start eating healthier.	Sarah has a bad habit of only eating candy and not eating healthy foods. This is a problem because she is not getting the nutrients she needs to stay healthy. Her dentist told her that she has cavities and needs to start eating more fruits and vegetables . . .	The dentist tells her she has 6 cavities.	Sarah has a bad habit of eating only candy. The dentist tells her she has 6 cavities. Sarah's bad habit of eating only candy has led to her having 6 cavities. The dentist told her this when she went for a check-up. Sarah needs to change her eating habits ...
Stan and his best friend have been friends for 8 years. They met when Stan was in first grade. They play games together, ate together, and watched movies together. Their favorite time of day is when Stan gets home from school. Stan's best friend is always waiting for him with a wagging tail.	Stan and his best friend have been friends for 8 years. They met in kindergarten and have been inseparable ever since. They do everything together, from playing in the park to watching movies at home. They even have a special handshake they use to greet each other. Stan's friend is always there for him, no matter what. . . .	Stan's best friend is always waiting for him with a wagging tail.	Stan and his best friend have been friends for 8 years. They met when Stan was walking home from school and his friend was wagging his tail and barking at him. Stan approached the dog and petted him, and they have been friends ever since. They spend a lot of time together, playing, going for walks, and just hanging out. . . .
Gina prepared food for the family. She cooked lasagna, grilled chicken and garlic bread. As soon as we started to eat, her mom became nauseous. Afterwards, everyone started to feel queasy. They eventually threw away the food and ordered Chinese food.	Gina prepared food for the family. She cooked lasagna. However, she got sick and threw up. The family ate the lasagna, but they all got sick and threw up. They went to the hospital and were diagnosed with food poisoning. The food was later found to be contaminated with *E. coli*.	Her mom became nauseous.	Gina prepared food for the family. She cooked lasagna. However, her mom became nauseous and threw up after eating it. Gina then realized that she had used spoiled ground beef for the lasagna. The rest of the family also became sick after eating it.

Table S1: Narrative examples of retrieval-augmented generation with compression as a model of hippocampal-neocortical interaction, using Mistral-7B-Instruct-v0.2 plus xRAG weights to simulate neocortex. The 'Original' column gives the story 'experienced' by the system. The 'Gist only' column gives the reconstruction from just the xRAG vector, i.e. when the xRAG vector is retrieved from the hippocampus and used to condition the generative network. The 'Stored detail' column lists the most surprising detail given the xRAG vector. Specifically, all phrases are extracted from the text, then the perplexity of each phrase given the xRAG vector is calculated. In other words, for each phrase, the model is prompted with the xRAG vector and the phrase in question. Then the phrases are ranked in descending order of perplexity, with the top phrase used as the 'Stored detail'. The text embedding (from which the xRAG vector is derived) and details for each memory are encoded together in the model hippocampus. The 'Gist + detail' column gives the reconstruction of the memory from this representation.

Dataset	Concreteness	Richness in detail	Specificity
HIPPOCORPUS	$r = 0.019$, $p < 0.001$	$r = 0.023$, $p < 0.001$	$r = 0.023$, $p < 0.001$
Simulated	$r = 0.279$, $p < 0.001$	$r = 0.350$, $p < 0.001$	$r = 0.240$, $p < 0.001$

Table S2: Pearson correlations between story length and each LLM-derived attribute.

Attribute	Recalled vs. Retold	Recalled vs. Imagined	Retold vs. Imagined
Concreteness	$t = 4.823$, $p < 0.001$	$t = 368.994$, $p < 0.001$	$t = 6.445$, $p < 0.001$
Richness in detail	$t = 5.467$, $p < 0.001$	$t = 445.861$, $p < 0.001$	$t = 7.770$, $p < 0.001$
Specificity	$t = 7.772$, $p < 0.001$	$t = 426.456$, $p < 0.001$	$t = 5.292$, $p < 0.001$

Table S3: The results of t-tests on length-residualised attribute scores for HIPPOCORPUS stories.

Attribute	Original vs. Encoded	Original vs. Consolidated	Encoded vs. Consolidated
Concreteness	$t = 3.796$, $p < 0.001$	$t = 5.850$, $p < 0.001$	$t = 2.430$, $p = 0.015$
Richness in detail	$t = 4.027$, $p < 0.001$	$t = 6.710$, $p < 0.001$	$t = 3.468$, $p < 0.001$
Specificity	$t = 4.688$, $p < 0.001$	$t = 6.581$, $p < 0.001$	$t = 2.211$, $p = 0.028$

Table S4: The results of paired t-tests on length-residualised attribute scores for simulated stories.

Type	Recalled story example
Original	One night two young men from Egulac went down to the river to hunt seals and while they were there it became foggy and calm. Then they heard war-cries, and they thought: "Maybe this is a war-party". They escaped to the shore, and hid behind a log. Now canoes came up, and they heard the noise of paddles, and saw one canoe coming up to them. There were five men in the canoe, and they said: . . .
Temperature of 0	One night two young men from Egulac went down to the river to hunt seals and while they were there it became foggy and calm. Then they heard war-cries, and they thought: "Maybe this is a war-party". They escaped to the shore, and hid behind a log. Now the war-party started, and they heard the noise of paddles, and saw one canoe coming up to them. There were five men in the canoe, and they said: ...
Temperature of 0.5	One night two young men from Egulac went out to the river to hunt seals and while they were there they were robbed by the men from Egulac. The men ran back to Egulac and hid behind a log. When they came out they discovered a canoe coming up to them and they jumped up and began to fight. The canoe came down to the ground and they began to wrestle. The crowd cheered for the young men of Egulac. . . .
Temperature of 1	One night two young men from Egulac went up to the river and they saw a canoe coming up to them. Then they heard war-cries, and they heard the sound of paddles, aand they heard war-cries, and they saw one canoe coming up to them. It was full speed, and when it arrived the men said: "This is a war-party. We have come to conquer you in a fight". When they reached the shore, they saw one canoe . . .

Table S5: Recalled stories for different temperatures, for a model trained for five epochs with the 'Sports' category of Wikipedia data (*62*) as the background data distribution, and prompted with 'One night two young men from Egulac'.

unit (*107*). In the sequential model, this could be adapted to store sequences as follows. Let us consider sequences represented as strings of characters or symbols, which can capture language, spatial trajectories, transitions in a graph, stimuli in sequential learning tasks, and more.

(*50*) give a unifying account of how neural network models of associative memory such as MCHNs operate, which helps to explain the extension to sequences. They develop the biologically plausible modern Hopfield network (*108*) into 'a general framework for understanding the operation of . . . memory networks as a sequence of three operations: similarity, separation, and projection', which they term 'universal Hopfield networks' ((*50*), Abstract). They observe that any of these models can be made asymmetric or

Inference template	Mean accuracy
{} EAST {} WEST {}	0.87
{} WEST {} EAST {}	0.82
{} NORTH {} SOUTH {}	0.81
{} SOUTH {} NORTH {}	0.87
{} EAST {} SOUTH {} WEST {} NORTH {}	0.71
{} SOUTH {} WEST {} NORTH {} EAST {}	0.78
{} WEST {} NORTH {} EAST {} SOUTH {}	0.8
{} NORTH {} EAST {} SOUTH {} WEST {}	0.81
{} EAST {} EAST {} NORTH {} WEST {} WEST {} SOUTH {}	0.79
{} NORTH {} NORTH {} WEST {} SOUTH {} SOUTH {} EAST {}	0.74

Table S6: Spatial task inference templates and their average accuracies.

Inference template	Mean accuracy
{} CHILD_OF {} PARENT_OF {}	0.81
{} PARENT_OF {} CHILD_OF {}	0.76
{} GRANDCHILD_OF {} GRANDPARENT_OF {}	0.8
{} GRANDPARENT_OF {} GRANDCHILD_OF {}	0.62
{} CHILD_OF {} CHILD_OF {} GRANDPARENT_OF {} SIBLING_OF {}	0.7
{} CHILD_OF {} SPOUSE_OF {} PARENT_OF {} SIBLING_OF {}	0.75
{} PARENT_OF {} SIBLING_OF {} CHILD_OF {} SPOUSE_OF {}	0.79
{} PARENT_OF {} PARENT_OF {} GRANDCHILD_OF {} SPOUSE_OF {}	0.77
{} CHILD_OF {} SPOUSE_OF {} CHILD_OF {} SPOUSE_OF {} GRANDPARENT_OF {} SIBLING_OF {}	0.75
{} GRANDPARENT_OF {} SIBLING_OF {} CHILD_OF {} SPOUSE_OF {} CHILD_OF {} SPOUSE_OF {}	0.72

Table S7: Family tree task inference templates and their average accuracies.

heteroassociative rather than autoassociative, if the projections from feature units to memory units capture the current state, but the projections from memory units back to feature units are different (for sequences, these projection weights would correspond to the 'next state'). Then one state retrieves the next, rather than each state retrieving a pattern-completed version of itself. That is, we want to store patterns X and 'next patterns' X_{next} instead of just patterns, and change the MCHN's update rule accordingly:

$$\xi_{next} = X_{next} softmax(\beta X^T \xi)$$

The modern continuous asymmetric Hopfield network (MCAHN) converts the MCHN to work for sequences, consistent with the framework above (*109*). However there is a problem when you try to apply this naively to certain sequences, as the MCAHN is by default a 'Markov chain' model of sequential memory. Whilst it works for sequences in which items are unique, it doesn't work for, e.g., sentences without modification, as it cannot accurately remember sequences with repeated items; see Figure 5a of (*110*). When an item is repeated in a sequence, (*110*) show that a MCAHN has trouble recalling the next item, instead producing a composite.

(*110*) introduce the temporal predictive coding network (tPC), a more complex network learnt by gradient descent, based on the intuition that higher layers minimize prediction error at lower layers (*111*). (A drawback of the tPC is that the model must be 'presented with the sequence for multiple epochs' until convergence, so it is no longer capable of one-shot learning, which arguably undermines its plausibility as a model of hippocampal encoding.) They show that the one-layer tPC also has the MCAHN's issue with repeated items, but suggest the two-layer tPC to resolve this. As (*110*) show in Figure 5d of their paper, the hidden layer represents an item's position in the sequence so the next item can be recalled correctly.

Alternatively, one could combine the concept of the modern asymmetric Hopfield network (with return projections from memory to feature units representing the next state) with a state representation that captures the history. This is based on previous work involving the role of temporal context in memory (*112, 113*). Consider the example of encoding sentences, i.e. sequences of characters. Suppose each state is a vector of length equal to the number of symbols, consisting of one at the index of the current symbol, plus the previous state multiplied by some decay factor. See Figure S1 for an example. This has the benefit of still being compatible with one-shot learning, although the memory capacity may not scale as well as the more complex predictive coding model approach. Initial testing suggests that this model can encode and retrieve sentences (although retrieval performance goes down with sentence length).

To show how the preceding 'context' is captured, consider the example of encoding sentences, i.e. sequences of characters. Each state can be represented by a vector of length equal to the number of symbols, with one at the index of the current symbol, plus the previous state multiplied by some decay factor. In other words, we want a vector representation $\mathbf{x_i}$ of the ith state which is a sum of the current symbol's vector $\mathbf{v}$ and activity from the previous state $\mathbf{x}_{i-1}$ multiplied by a decay rate λ (noting that the states are then normalised to unit length):

$$\mathbf{x}_i = \frac{\mathbf{v}_i + \lambda \cdot \mathbf{x}_{i-1}}{\|\mathbf{x}_i\|}$$

As with all modern Hopfield network (*49*) variants, the inverse temperature, β, is a key parameter determining the network's behaviour. In this case a high value of β would be desirable to avoid composite states. The softmax function is given below; in the limit of high β, the output of this function becomes a vector with one at the index of the maximum and value and zero elsewhere:

$$softmax(z_i) = \frac{e^{\beta z_i}}{\sum_{j=1}^{K} e^{\beta z_j}}$$

Whilst this simple model helps justify the assumption that the hippocampus can store sequences associatively, it clearly has some limitations. For example, it suggests that when a state from a certain point in an encoded sequence triggers recall, states preceding that point cannot be recalled (only the subsequent states). This is clearly not the case, so a model of sequence memory that can only retrieve states in one temporal direction is unsatisfactory. Also, if states involve vectors of continuous rather than binary variables, the decaying activity from previous states could be hard to distinguish from the current state, so a more complex solution could be required. In addition, the sequences consist of a single symbol at any moment in time, but other sequences like frames in a video consist of complex representations at each time step. The Hopfield network variant could perhaps have a combination of autoassociative and heteroassociative connectivity to account for this, with autoassociative connections reconstructing the current time step, and heteroassociative connections activating the next (see (*81*)).

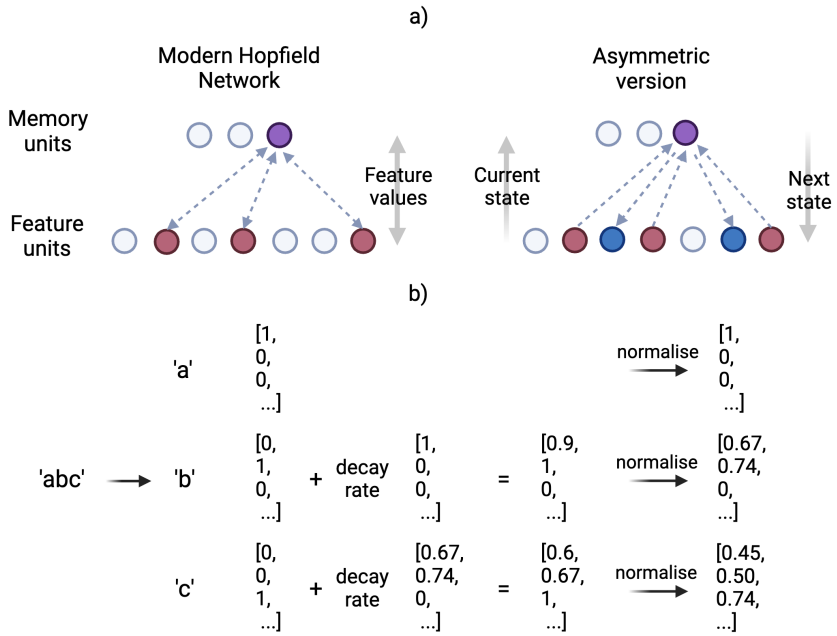

Figure S1: Using asymmetric modern Hopfield networks to model the hippocampus. a) In asymmetric versions of modern Hopfield networks, the weights between feature and memory units differ, as visualised from the perspective of (*107*). b) A sequence of arbitrary symbols, e.g. 'abc', is represented as a sequence of vectors over the feature units, where each vector is the sum of the current state plus the decay rate (here set to 0.9) times the previous state. For each state, the return connections from memory units to feature units encode the next state.