Trends in Cognitive Sciences

CellPress

Review

Towards large language models with human-like episodic memory

Cody V. Dong [1], Qihong Lu [2], Kenneth A. Norman [1,3,*], and Sebastian Michelmann [4,*]

Cognitive neuroscience research has made tremendous progress over the past decade in addressing how episodic memory (EM; memory for unique past experiences) supports our ability to understand real-world events. Despite this progress, we still lack a computational modeling framework that is able to generate precise predictions regarding how EM will be used when processing high-dimensional naturalistic stimuli. Recent work in machine learning that augments large language models (LLMs) with external memory could potentially accomplish this, but current popular approaches are misaligned with human memory in various ways. This review surveys these differences, suggests criteria for benchmark tasks to promote alignment with human EM, and ends with potential methods to evaluate predictions from memory-augmented models using neuroimaging techniques.

Highlights

A fundamental question in memory research is how episodic memory (EM) supports understanding of real-world events.

Progress in this area requires computational models that can help us to make sense of the complex ways in which EM interacts with semantic memory and working memory when processing naturalistic stimuli.

Memory-augmented large language models (MA-LLMs) could potentially play this role because they possess rich semantic knowledge that allows them to respond appropriately to naturalistic inputs, and they also have an external memory system akin to EM.

This review evaluates how well these models capture key properties of human EM (relating to dynamic memory updating, event segmentation, selective encoding and retrieval, temporal contiguity, and competition at retrieval) and describes ways in which they could be better aligned with human data, thereby improving their suitability as cognitive models.

Understanding the role of episodic memory in real-world prediction

Most of our knowledge about episodic memory (**EM**; memory for unique past experiences, see Glossary) has come from laboratory studies that directly ask participants about the past. However, over the years there has been a growing appreciation of the myriad ways in which EM is used outside of this question-answering scenario. Notably, laboratory studies of incidental learning of item sequences [1,2] and eye-movement control [3–5] have shown that, when stimuli repeat, we can draw upon our EMs of prior encounters with these stimuli to generate predictions about what will happen next. Another important insight is that, in real-world scenarios, the contribution of EM to prediction is inextricably intertwined with the contributions of **semantic memory** and **working memory**. While EM supports rapid memorization of rich, contextualized (who, what, where) details from unique, personally experienced events, semantic memory gradually integrates across multiple events to represent the shared structure of those events [6], and working memory supports the maintenance and manipulation of information over short timescales. Consider a situation where you are at a restaurant with a friend: your prediction about what your friend will order will depend on interactions between semantic memory (what they usually order), EM (your recollection of what they ordered last time), and working memory (they just told you they want to try something different).

Cognitive neuroscience research has made tremendous progress over the past decade in characterizing how interactions between memory systems support real-world prediction. This progress has been driven by studies that have measured brain activity while participants process naturalistic stimuli such as movies and verbal narratives (e.g., [7]). Some of these studies have asked participants to explicitly recall naturalistic materials (e.g., TV episodes [8]), whereas others have looked at activity related to EM encoding and/or retrieval during comprehension of narratives [9–15]. These studies have revealed several important properties of how EM works in the context of naturalistic stimuli, for example the role of event boundaries in shaping EM encoding and retrieval (discussed below).

[1]Department of Psychology, Princeton University, Princeton, NJ 08540, USA
[2]Zuckerman Institute, Columbia University, New York, NY 10027, USA
[3]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA
[4]Department of Psychology, New York University, New York, NY 10003, USA

*Correspondence:
knorman@princeton.edu (K.A. Norman) and s.michelmann@nyu.edu (S. Michelmann).

© 2025 Elsevier Ltd. All rights reserved, including those for text and data mining, AI training, and similar technologies.

Despite this progress, the complexity of these interactions between memory systems poses a strong challenge for understanding real-world memory. Computational models can play a useful role in taming this complexity. Toward this end, researchers have started to use simple neural network models to explore how interactions between memory systems can support adaptive behavior (e.g., [16,17]); however, these small-scale models lack the rich semantic knowledge that people have, which limits their ability to make predictions about specific real-world situations. We describe below how **memory-augmented LLMs (MA-LLMs)** may provide a solution to this problem – like other types of **LLMs**, MA-LLMs possess rich, context-sensitive semantic knowledge stored in their weights (akin to human semantic memory) and they can maintain and manipulate recently presented information (akin to human working memory), allowing them to process naturalistic inputs. In addition to these core LLM features, MA-LLMs possess an additional memory system that can support recall of unique events (akin to human EM; for reviews, see [18,19]). To a first approximation, MA-LLMs appear to have the right 'pieces' to serve as cognitive models of how memory systems interact to support real-world prediction. However, given that MA-LLMs were not designed with the goal of fitting human data, it stands to reason that the properties of MA-LLMs may be misaligned with those of human EM [20]. In this review we compare the properties of extant MA-LLMs to human EM, with the primary goal of identifying ways to make MA-LLMs more human-like, so they will be more effective as cognitive models. In addition, to the extent that human memory includes useful features that are not presently incorporated in MA-LLMs, improving alignment to human EM may further serve to advance artificial intelligence (AI).

The role of external memory in extending LLMs

LLMs are neural networks that are trained to predict missing words over large amounts of training text; in so doing, LLMs acquire a wealth of semantic knowledge based on how naturalistic text typically unfolds [21]. Standard (non-memory-augmented) LLMs excel at capturing the statistics of human language (e.g., [22]; for a review, see [23]), but memorization of once-presented information across long timescales (as tested, e.g., in [24]) poses more of a challenge. In this section we explain why this is the case and how this has led to the development of MA-LLMs.

One way that LLMs learn is by incrementally adjusting their synaptic weights over the course of training. Because of the incremental nature of training, standard LLMs are only likely to memorize information that is repeated across the training set [25–27]. Although training with a higher learning rate could lead to better acquisition of information that is only presented once, this comes at the risk of catastrophic interference (i.e., rapid forgetting of prior knowledge [28,29]).

In addition to storing knowledge in their weights, the transformer architecture used by most modern LLMs [30] allows them to maintain a window of text called the **context window** (for alternatives to the transformer architecture, see [31,32]). Through a process known as self-attention (Box 1), transformer-based LLMs are able to use the knowledge stored in their weights to flexibly combine information in the context window, yielding a context-sensitive representation of the current item that supports more accurate prediction [22]. Representing information 'in context' has been shown to support improved generalization, beyond what is observed based on incremental weight adjustment alone [33]. Importantly, the context window provides a way of remembering information that was only presented once, provided that it fits in the window. In recent years, researchers have sought to vastly expand the size of the context window (e.g., 128k tokens for GPT-4-128L), effectively pushing its function outside of the domain of working memory (limited-time storage of recent information) into the domain traditionally associated with EM (storage of much larger numbers of memories across much longer timescales; see Box 1 for further discussion of how in-context memory relates to working memory and EM).

Glossary

Context window: a window of contiguous text inputs that an LLM has access to when interpreting the current word.
Episodic memory (EM): a memory system that binds together the features of a unique experience (episode) in a lasting way such that the contents of that episode (i.e., what happened) can be retrieved later together with contextual features (e.g., where and when the episode took place). In humans, episodic memory depends on neural processes in the hippocampal formation and the surrounding medial temporal lobe.
Large language model (LLM): a neural network trained on a large body of text to predict held-out words from text passages.
Memory-augmented large language model (MA-LLM): an LLM augmented with a memory store that allows the LLM to retrieve information from outside the context window.
Semantic memory: a memory system that stores knowledge acquired by integrating across multiple experiences, where this integration makes it possible to identify common, generalizable structure. In humans, semantic memory largely relies on neural processes in neocortex.
Working memory: a memory system that supports the maintenance and manipulation of information over short timescales. Working memory is limited in capacity but can dynamically interact with other memory systems; for instance, when episodic memories are retrieved, they are loaded into working memory. In humans, working memory has been linked to neural processes in prefrontal cortex, but other brain regions have also been implicated in working memory processes.

Box 1. Relating in-context memory to human memory systems

The transformer architecture used by most modern LLMs relies on a computation called self-attention [30,109] whereby each token (stimulus) processed by the network generates three distinct vectors: a query, a key, and a value (Figure IA). The query generated by the current input is compared against the keys associated with stimuli presented earlier in the context window – these query–key match scores are used to compute a weighted combination of the values associated with the keys (i.e., the better the query matches the key, the more the value associated with that key is included in the weighted average). The process is called self-attention because the model 'attends' to relevant parts of the context to interpret its current input. From a cognitive modeling perspective, if we assume that the contents of the context window are active in mind (because they were recently presented), self-attention can be viewed as an implementation of working memory manipulation (i.e., flexibly recombining current thoughts). In long-context models, it is not plausible to assert that the entire context window is actively represented because the size of the context window in these models exceeds the known (limited) capacity of human working memory. One possible response is to argue that the self-attention computation does 'double duty' as an implementation of working memory (when applied to proximal parts of the window) and also something more like EM (when applied to distal parts of the window). That is, if one views the keys and values in the distal part of the context window as being latent (i.e., not currently active) memories, the self-attention computation can be viewed as an implementation of retrieval from long-term storage, whereby stored memory traces are made active according to how well they match the current query (Figure IB). Indeed, this query–key–value structure is a central part of both older 'global match' models of EM [110] and more recent theoretical models of long-term memory [111], and multiple recent papers have discussed how the hippocampus could implement self-attention (e.g., [112,113]). Using the exact same self-attention computation to model working memory and EM has simplicity in its favor; however, this approach lacks flexibility relative to the MA-LLMs discussed elsewhere in the paper, which use a separate system for EM, and thus are better positioned to account for the many functional dissociations that have been observed between working memory and EM in humans (e.g., [114]).

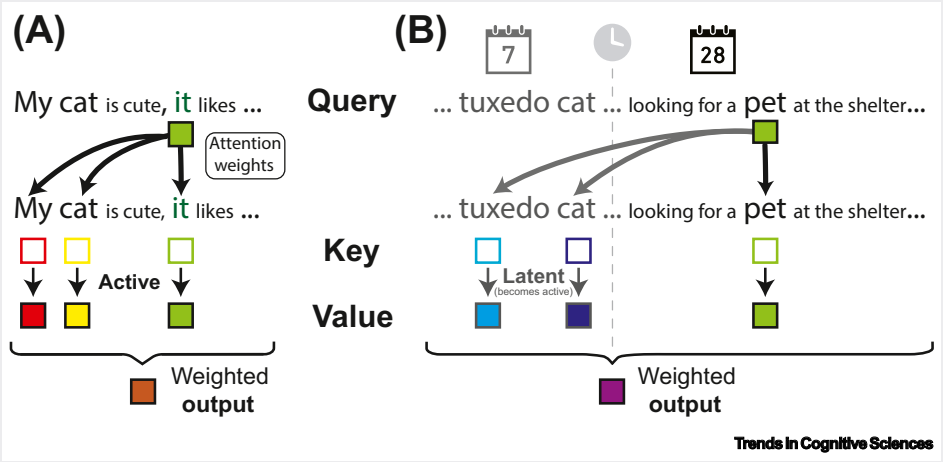


Figure I. Self-attention across the context window.

However, this long-context approach has faced practical issues relating to the high computational costs of scaling the self-attention computation as the context window increases in size [34].

The aforementioned limitations of learning by adjusting weights and of long-context models raise the question of whether there is a better way to support lasting one-shot memorization. In this regard we can look to humans, who solve this by having a separate system for EM (hippocampus). According to the influential complementary learning systems framework, the hippocampus uses relatively distinct internal representations that allow it to do one-shot memorization without suffering catastrophic interference, whereas neocortex uses more overlapping representations that facilitate the gradual acquisition of semantic knowledge across multiple experiences [35]. LLMs can also use this solution by adding a memory system that rapidly stores information in a latent and lasting form (like the hippocampus in humans) and is separate from both the context window (which can be limited in size, like working memory in humans) and the weights of the main LLM

(which are updated incrementally, like neocortex in humans). In this architecture, latent information in EM can be rendered active by retrieving it into the context window, where the retrieved information is then processed by the main LLM. This arrangement preserves the separation between EM and working memory that is present in humans, and stands in contrast to long-context models where the context window serves the functions associated with both EM and working memory. Given the strong computational benefits of using a separate EM system (one-shot memorization without catastrophic interference, while avoiding the computational costs associated with long-context models), recent machine learning research has adopted this solution as a natural way to extend LLMs [18,19,36].

Memory architecture in MA-LLMs

Memory-augmented LLMs take many forms (see [18,19,37] for reviews), but the main premise is to allow LLMs to access and use relevant stored pieces of information from outside the context window during ongoing text generation. Retrieval-augmented generation (RAG) is a popular MA-LLM variant where relevant information from an external datastore is retrieved and subsequently leveraged by the LLM to improve response generation (Box 2).

An important property of RAG is that information can be directly added to its datastore without processing it through the main LLM. By contrast, other forms of MA-LLMs require the information to be processed through the main LLM in order to be stored. As described in Box 1, LLMs that use the transformer architecture [30] generate keys and values for each item (e.g., word) they process – some MA-LLMs [38,39] simply take chunks of keys and values that were generated when an item was 'in context' and store them in EM. Later, those stored chunks of keys and values are retrieved according to some retrieval rule; once retrieved, the keys are treated similarly to the keys from the current context (i.e., the query generated by the current token is matched to the retrieved keys as well as to the keys corresponding to tokens presently in the context window). Essentially, the difference between RAG and these other models corresponds to the difference between storing the verbatim stimulus (what RAG does) versus storing an internal representation of the stimulus (the keys and values). The latter approach aligns better with human EM, insofar as humans store internal representations rather than verbatim input.

Comparing human EM to memory in MA-LLMs

External memory in MA-LLMs is aligned with human EM in several important ways: storage is latent and effectively unbounded; one exposure is sufficient for storage, and EMs are blended with semantic knowledge to generate predictions [40] – in MA-LLMs, this is accomplished by feeding retrieved EMs into the core LLM (where semantic knowledge resides), which then generates the

Box 2. Retrieval-augmented generation

RAG refers to a class of MA-LLM models that are composed of a datastore, a retriever, and an LLM text generator [18,37]. The datastore is often a large set of external documents that are useful for the downstream task (usually question answering). The retriever can vary depending on the model, but popular information retrievers include BM-25 (a sparse retriever that identifies relevant memory chunks based on word co-occurrences [115]) and dense retrievers such as Contriever [116] that map retrieval queries and text documents to a latent semantic vector space for later retrieval (where retrieval is a function of vector similarity between queries and text documents). The top k relevant documents are retrieved (where k is usually greater than 1) and these retrieved documents are fed into the LLM; a common approach ('in-context RAG') is to simply concatenate the retrieved documents to the input context of the model for text generation [72]. With this approach, the retrieved information is treated in the same way as other, recently presented information (akin to how humans load retrieved EMs into working memory). In-context RAG is therefore an easily adaptable and deployable method that uses off-the-shelf information retrieval systems and off-the-shelf language models without need for any additional pretraining/fine-tuning or modifications to the underlying retriever and LLM architectures. Other forms of RAG train retrievers and LLM components jointly [63,64] or sequentially [65], which could potentially allow the models to do a better job of using the retrieved information.

model's prediction. In this section we describe other foundational properties of human EM and evaluate how well they align with the properties of external memory in MA-LLMs.

Dynamic memory updating

In humans, new experiences leave lasting EM traces – we are constantly adding memories. Importantly, a large body of evidence suggests that memories can be altered after having been added (Box 3). Episodic memories are instantiated in the brain as sparse distributed patterns [41] that are stored by graded adjustments to synaptic weights; this arrangement allows nuanced updating of both the strength and content of memories. Although MA-LLMs are capable of adding new EMs, it is not common to update these memories once they have been added. Enhancing memory-augmented models by allowing models to forget irrelevant (or outdated) EMs and change stored EMs is an active area of research [42–44]. The Titans model [45] is particularly promising in this regard because it learns to store new EMs (key–value pairs) in a weight matrix that is continually updated as new information comes in (see also [46]), as opposed to storing them in an all-or-none fashion.

Importantly, in addition to dynamic updates to EM itself, EM is used to update semantic memory through a process of consolidation; in the brain, this process involves replay of hippocampal EMs which drive learning in the neocortex as it extracts shared structure across replayed EMs [35,40,47–49]. Although replay has been used in modern neural networks to support updating of knowledge stored in the network's weights [50,51], it is still not common in MA-LLMs to replay stored EMs to update weights in the main LLM (although there are some exceptions; e.g., [52,53]).

Event segmentation

People segment continuous real-world experience into discrete events (e.g., a conversation or a work meeting); this process of event segmentation is automatic and people produce reliable judgments for when one meaningful event ends and another begins [54]. These moments of transition between events (event boundaries) are reflected in the brain as shifts in stable patterns of neural activity throughout the neocortex [9,55]. Current mathematical models of event segmentation posit that event boundaries correspond to moments when one's inference about the underlying (latent) cause of one's experiences changes [56,57]; these moments often (but not always) correspond to spikes in prediction error [58–60]. Importantly, event boundaries have been shown

Box 3. The malleability of human memory

Evidence suggests that human memories can be modified in a wide range of ways after they are initially formed. Some of these alterations involve adjusting the strength of memories, thereby making them more or less accessible. One theory posits that these changes can occur as a result of competition during retrieval, such that the winning (best-matching) memory is strengthened and other memories that come to mind to a moderate degree (but not as much as the winning memory) are weakened [117] – these changes can be seen as sculpting the landscape of memory storage to make it less likely that people will confuse these memories in the future. In addition to altering the strength of memories, studies of memory reconsolidation suggest that it is possible to retrospectively adjust the content of memories (e.g., updating them based on new information) or even delete them outright [118] – these changes can be useful in keeping the memory system in sync with a changing world [119]. A unifying feature of all of these ideas about memory alterations is the central role of reactivation in memory updating (i.e., memories are only updated in strength or content to the extent that they are activated during retrieval). Another important feature of human memory is that alterations can take place in both the hippocampal representations associated with memories (e.g., competition between memories can make these hippocampal representations more distinct [117,120–122]) as well as in the content (represented in neocortex) associated with the memories [123]. A recent paper [112] argues that this distinction between hippocampal and neocortical representations corresponds to the key–value distinction described earlier, where the hippocampus stores keys that are optimized for discriminability; these hippocampal keys support retrieval via associative connections to value representations stored in neocortex. In this framework, altering the hippocampal keys makes it possible to change the retrievability of memories (e.g., to minimize competition) without affecting the content of the memories once they are retrieved [112].

to play a fundamental role in organizing EM. For example, a recent study showed that people can 'skip ahead' to the next event boundary when trying to sequentially scan through their memories ([61]; for a review of how event boundaries influence memory, see [62]).

In contrast to the dynamic event segmentation used by humans, MA-LLMs generally store their memories in fixed-size chunks [63–65] (Figure 1, Key Figure). A notable exception is the EM-LLM model [66], which segments the ongoing stream of tokens into events based on surprise and then stores these event chunks in EM. Results from this model show a strong functional benefit of event segmentation; to explain this benefit, the authors argue that dividing the timeseries of inputs into chunks of related information helps to minimize retrieval of irrelevant information.

Selective encoding and retrieval

In humans, encoding and retrieval occur selectively (i.e., they occur more strongly in some circumstances than others). For example, one study [13] found that participants showed stronger

Key figure

Contrasting episodic memory (EM) in memory-augmented large language models (MA-LLMs) and humans

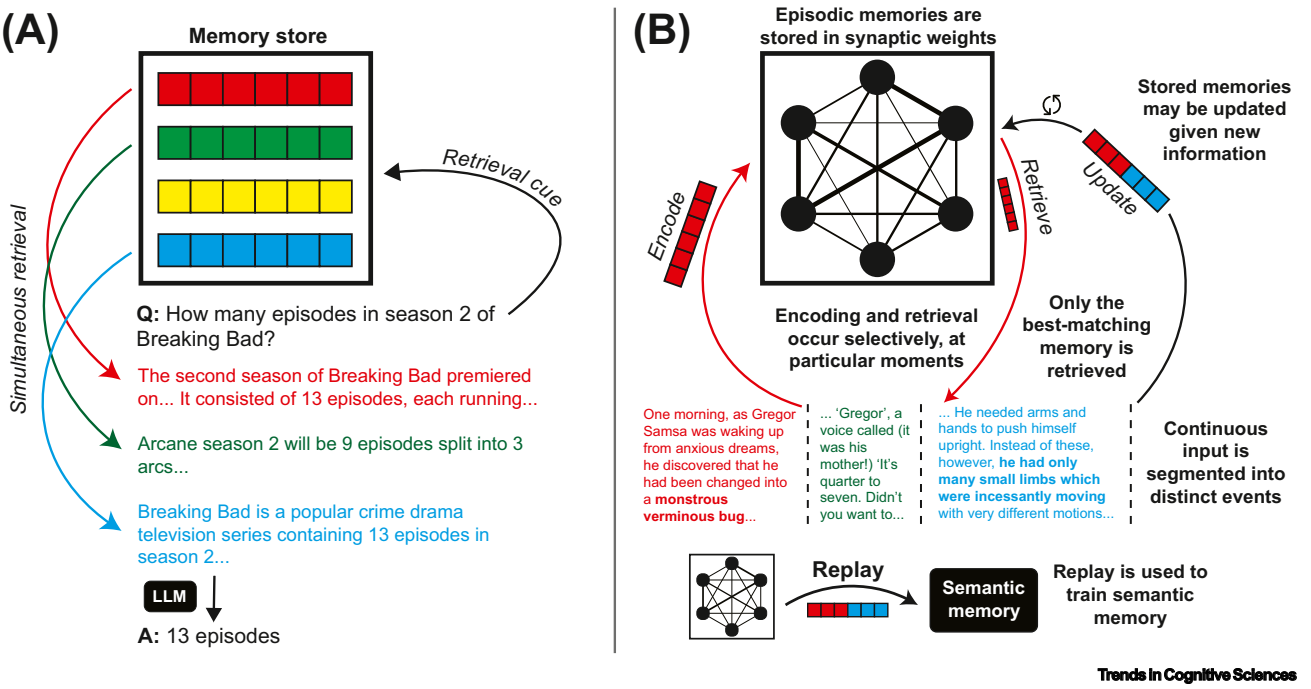


Figure 1. (A) Illustration of a common approach to augment LLMs with memory retrieval. Typical aspects of MA-LLMs include memories composed of uniform-length chunks that are not updated after initial storage, and simultaneous retrieval of multiple memories at once [64,72]. (B) By contrast, human memory is dynamic; episodic memories are constantly being added, forgotten, and updated [123] through nuanced adjustments to synaptic weights, and replay of episodic memories is used to train semantic memory through a process of consolidation. Continuous input is segmented into distinct events [124], and encoding and retrieval occur selectively – EM retrieval is more likely to occur when there are 'gaps' in understanding to fill [13], and some moments, such as boundaries between events, tend to be associated with stronger encoding and retrieval than others [10,15,67]. Memory retrieval is competitive, meaning that (most of the time) only the single best-matching memory is retrieved (e.g., [125]). Another key property, not pictured here, is that EM for the content of events is stored together with a representation of temporal context (what came before the event; e.g., [77]) that supports temporal contiguity effects in retrieval (i.e., successive recalls tend to come from nearby timepoints).

neural correlates of EM retrieval when there were momentary 'gaps' in their understanding of a movie (because the first part of the movie was viewed 24 h previously) versus when the exact same movie was viewed continuously. A recent paper [16] provided a normative account for this finding, arguing that EM retrieval was suppressed in the continuous-viewing condition because the information needed to make accurate predictions was already active in working memory, and consequently the benefits of EM retrieval were small relative to the potential risks of EM retrieval (in particular, retrieving the wrong memory, which could lead to inaccurate predictions). Relatedly, event boundaries have been found to be important points for EM retrieval [15,67]. This fits with the idea that, within a typical event (e.g., a familiar train ride), it is possible to predict what will happen next based on semantic memory, so the incremental benefits of EM retrieval are low [16]; however, when the event ends, there is a spike in uncertainty (when the train arrives in the city, there are many places you could go) [68,69] that can potentially be resolved by retrieving a relevant EM (e.g., remembering that, the last time you went into the city with this friend, they wanted to go ice skating).

Neural studies have also found that event boundaries are important points for encoding [9,15,70,71]. This result can be explained normatively, based on the idea that memories encoded at the end of an event have more complete information than memories encoded midway through; encoding mid-event can therefore lead to redundant memory 'clutter' that makes it harder to retrieve the most informative memory [16]. This prediction was supported by subsequent empirical work demonstrating that, during novel movie watching, neural correlates of mid-event encoding were associated with worse subsequent recall of those events [10].

In contrast to the selectivity exhibited by humans, MA-LLMs are typically unselective in how they encode and retrieve EMs. Models that store key–value pairs in EM typically store all of these pairs (e.g., [38,66]), and retrieval from EM typically occurs in a rote fashion (e.g., once every k tokens [72]). One exception is FLARE [73], a model that selectively triggers memory retrieval only when its prediction of the upcoming sentence without memory retrieval (i.e., using semantic memory) is uncertain. The authors observe a strong functional benefit of this selectivity, and their explanation of these benefits aligns closely with the aforementioned normative analysis of retrieval selectivity [16]: suppressing retrieval when the model is certain reduces retrieval of irrelevant memories that (if retrieved) would degrade prediction.

Temporal contiguity

Humans show temporal contiguity effects (i.e., successive recalls tend to come from nearby timepoints [74]). This has been explained using models that compute a drifting temporal context representation composed of a running average of recently presented items [75–77]. This temporal context representation is bound to item representations in EM. When an item is retrieved, the associated temporal context representation is also retrieved; cueing with this just-retrieved temporal context representation favors retrieval of items that were studied close in time to the just-retrieved item. Importantly, these contiguity effects have been shown to be scale-invariant (i.e., they apply across multiple timescales [75]).

MA-LLMs generally do not show temporal contiguity in the EMs they retrieve. One exception is the EM-LLM model [66], which directly imposes a simple form of local (not scale-invariant) contiguity on recalls from EM (see also [78]). Notably, a recent paper [79] found that standard (non-memory-augmented) LLMs will, in some circumstances, learn on their own to show temporal contiguity effects in how they process stimuli within the context window; an important question for future research is whether similar 'emergent temporal contiguity' effects could arise in how MA-LLMs retrieve EMs.

Competition at retrieval

Behavioral data suggest that humans show a strong amount of competition at retrieval, such that only the single best-matching memory comes to mind (or else no memory is retrieved). This limit on retrieval bandwidth is illustrated by classic work on cue overload [80] and the fan effect [81], showing that learning multiple associates to a cue impairs retrieval. Importantly, recent studies [82,83] have shown that activation can spread to a larger number of associates outside conscious processing, suggesting that the narrow retrieval bandwidth observed during conscious memory retrieval is likely to be a functional optimization rather than an architectural limitation of the memory system. Speculatively, this type of bandwidth limitation could serve as a useful constraint in situations where the number of relevant EMs is typically small (1 or 0). Like people, MA-LLMs limit the number of retrieved EMs; however, unlike people, MA-LLMs are typically parameterized to allow retrieval of more than one EM at a time [64]. To some extent, MA-LLMs – like people – can mitigate the costs of retrieving irrelevant EMs through some type of post-retrieval filtering [84–86], but it is unlikely that this approach will be completely effective (especially when incorrect information is similar to, and thus confusable with, the sought-after information [87]).

Aligning MA-LLM benchmarks with real-world memory demands

In machine learning, model development is heavily motivated by existing benchmark tasks. We focus here on how the properties of these common benchmark tasks differ from the demands of real-world memory use, and how this may account for some of the architectural differences described in the previous section.

Existing applications of MA-LLMs have mostly focused on question-answering (QA) tasks that measure the ability of a model to retrieve relevant information from its datastore (e.g., [27,88]). In QA tasks, the training and evaluation sets usually contain well-defined queries and answers, making it explicit when models should retrieve information (i.e., when they are prompted with the question) and also what information is relevant (i.e., the questions contain sufficient diagnostic information to uniquely implicate the sought-after memory). In addition, the correct answer is almost always stored in memory, often in multiple places.

In contrast to these machine learning benchmarks that use well-specified prompts, information in real life is continuously revealed and people need to decide when to retrieve (if at all) from memory. In many situations the information revealed thus far may not be sufficient to specify which memories are relevant; and there may not even be a relevant stored memory. Consequently, retrieval in real-world circumstances can carry a substantial risk (if a person or model tries to retrieve with an insufficiently specific cue, they could end up retrieving the wrong memory, and responding with the best-matching memory can lead to errors if there is no relevant memory in the first place).

These challenges of real-world human EM use may account for many of the properties of human EM reviewed above. For example, selective retrieval policies can mitigate the risks of retrieving based on an insufficiently specific cue. Retrieval timing can be framed as a speed–accuracy tradeoff: waiting to observe more information before retrieving can help to specify which memories, if any, are relevant – but, at the same time, waiting too long also has costs (because you deprive yourself of the benefits of retrieval) [16]. Learning a policy for when to retrieve that balances the costs and benefits of waiting can lead to improved performance. Selectivity at encoding may also help: less encoding leads to less clutter, and thus less potential to retrieve the wrong memory. Furthermore, competition at retrieval can help if relevant memories are scarce and many similar lures are present (by limiting the number of irrelevant and confusable memories that come to mind), and temporal contiguity can help to keep retrieval focused on details from relevant events [89].

The above discussion implies that use of benchmark tasks that do a better job of capturing real-world memory task demands will foster the development of more human-like MA-LLMs – an example of a task that better instantiates these properties is provided in Box 4. Other new approaches to benchmarking EM are also promising: for example, one new benchmark probes temporal context memory by asking the model to identify which of two passages from a book was presented first [90], and another tests the ability of models to accurately bind together and recall details from episodic events contained in artificially generated book chapters [91]; a further approach is to evaluate how well MA-LLM agents perform complex tasks that require memory (e.g., web shopping, multi-session conversations [92–94]).

Evaluating memory-augmented LLMs as cognitive models

The use of human-like tasks (such as that shown in Box 4) can encourage the development of human-like models, but – ultimately – the measure of whether a model is 'human-like' is how well it can explain experimental data from humans. One way to evaluate this is to use an encoding

Box 4. Creating a task that encompasses real-life memory challenges

Human memory in real life faces several challenges that promote the use of selective encoding and retrieval policies. To capture these challenges, a task should have the following properties.

(i) Information is presented continuously, such that the models need to decide when it is best to store and retrieve memories depending on what has been shown thus far.
(ii) It is not safe to assume that relevant information has been stored in memory.
(iii) When relevant information is presented, it only appears once (mirroring how, in the real world, information is often only presented once), and it is similar to other, irrelevant information, creating the potential for confusion.

Figure I shows an example of a task that would fulfill these criteria. During the encoding phase, a model is given a single exposure to (each of) a very large number of TV scripts for shows that came out after the weights of the model were trained (thereby ensuring that the model is seeing the scripts for the first time). During the task phase, the model is shown summaries of TV episodes, one sentence at a time; some of these episodes will have been included in the encoding phase but others will be new (not shown during the encoding phase or during the initial model training). The model has the option after each sentence to say 'continue' (i.e., view another sentence of the summary) or else it can 'take over' and summarize the rest of the episode. After taking over, points are awarded for recalling accurate details (not already presented in the summary) and deducted for recalling inaccurate details; if the model refrains from taking over then no points are awarded or deducted. The fact that summaries are presented incrementally, coupled with the existence of many similar scripts in memory, creates a strong risk of retrieving wrong information; high-scoring models will need to balance the need to wait for more information (to ensure that the correct memory is adequately specified) against the opportunity costs of waiting. Likewise, high-scoring models will need to avoid responding based on memory when no relevant memory was stored. To focus on the specific contributions of EM as opposed to prior semantic knowledge found in the base LLM, the performance of the memory-augmented model can be compared to the performance of a model that was not shown any scripts during the encoding phase.

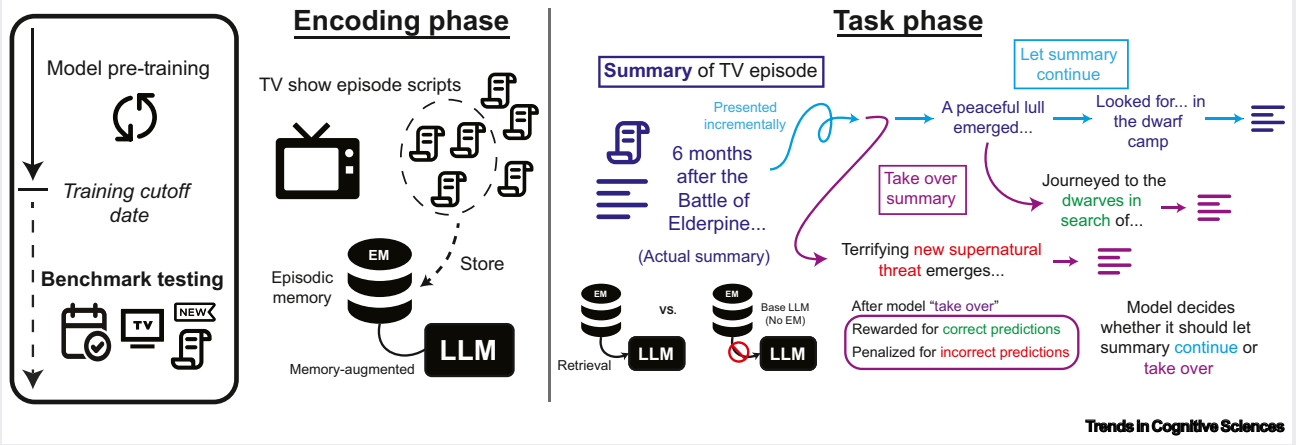


Figure I. An example of a task that reflects real-world memory challenges.

model approach, where MA-LLMs are evaluated based on how well their internal representations (embeddings) predict ongoing brain activity (Figure 2).

Another approach to evaluating MA-LLMs (apart from using encoding models) is to look at how well they can predict the timing of EM encoding and retrieval. Cognitive neuroscience researchers have identified several different neural correlates of when encoding and retrieval are taking place. For example, with invasive electrocorticography, it is possible to use time-lagged mutual information analysis to detect moments of information flow from neocortex to hippocampus (indicating EM encoding) and information flow from hippocampus to neocortex (indicating EM retrieval) [15]; with fMRI, inter-subject correlations in activity timeseries between hippocampus and default-mode cortical regions during movie watching and story listening have been linked to episodic retrieval [12,13], and increases in hippocampal activity have been linked to episodic encoding [9,70,71]. One could use an MA-LLM to predict the specific moments in a text narrative when EM encoding and/or retrieval should take place (e.g., at event boundaries), and then use the aforementioned neural measures to test these predictions.

A third approach to evaluating MA-LLMs is to see how well different MA-LLMs predict behavior. For example, if two individuals engage in repeated bouts of conversation (separated in time from one another), the present conversation will likely be influenced by memories from earlier conversations; different techniques for memory augmentation can therefore be evaluated on how well

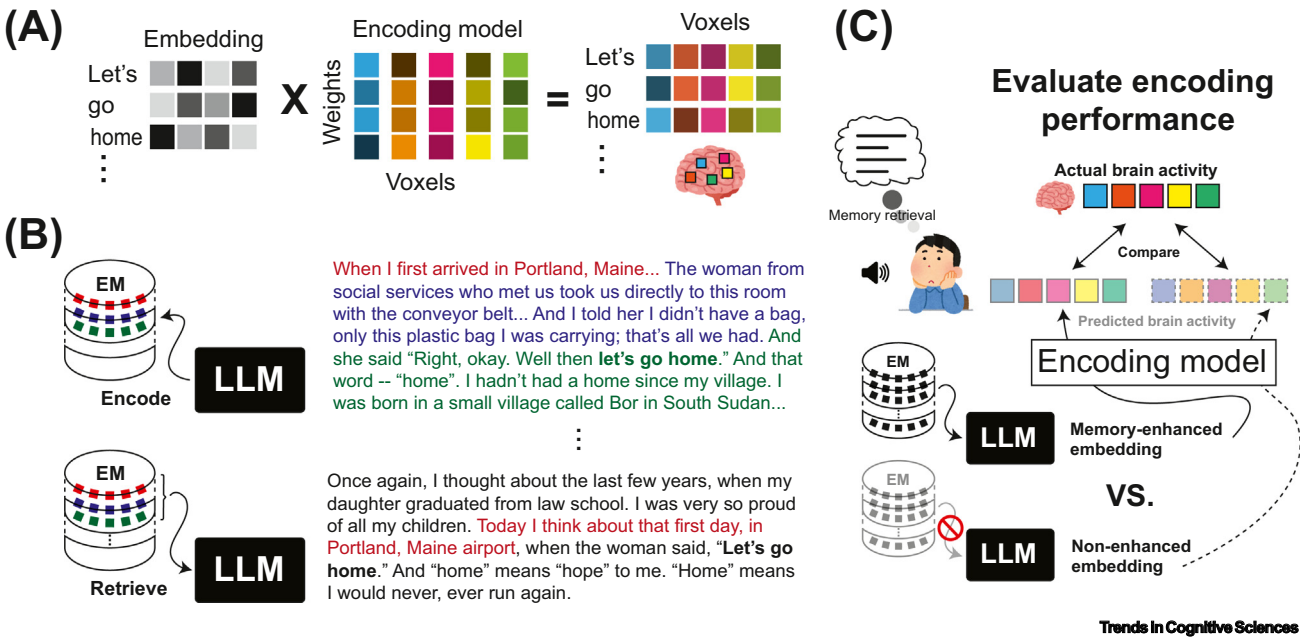


Figure 2. Testing memory-augmented large language models (MA-LLMs) using encoding models. One way to evaluate the fit of LLMs to brain data is to use an encoding model approach. This approach involves learning a mapping from features derived from an LLM (e.g., embeddings from one of the LLM's hidden layers) to brain activity (e.g., fMRI voxels; A); once learned, this mapping can be used to predict held-out data, and the fit to this held-out data can be compared across different encoding models that make use of different feature sets [126,127]. This encoding model approach has been used to answer a wide range of cognitive questions, including the extent to which a particular brain region is predicting upcoming stimuli [22,128] and the size of the temporal window that a given brain region is integrating over [129,130], and it can easily be adapted to quantify the benefits of memory augmentation (vs. no memory augmentation). (B) Illustration of an MA-LLM that stores chunks of text and then retrieves those chunks in response to a later part of the narrative. Predictions of brain activity generated using embeddings from the MA-LLM can be compared to predictions of brain activity generated using embeddings from a closely matched, non-memory-augmented LLM (C); this approach could also be used to address more fine-grained questions (e.g., arbitrating between two distinct MA-LLMs that make different predictions about what information will be retrieved). Abbreviation: EM, episodic memory.

they are able to leverage these episodic memories (in tandem with the current conversational context) to predict what the participants will say next.

Concluding remarks

In this review we have focused on identifying fundamental architectural properties of human EM that are not yet widely incorporated in MA-LLMs; we have discussed how to construct tasks that more closely approximate the challenges faced by human EM, and we have described general approaches to testing the fit between MA-LLMs and data from humans. Bringing MA-LLMs into closer alignment with human EM could substantially advance the cognitive modeling of memory, making it possible – for the first time – to predict human EM use in response to a novel naturalistic stimulus. There are also potential benefits for AI in better aligning MA-LLMs with human EM: as discussed earlier in the paper, models that incorporate key properties of human EM such as retrieval selectivity [73] and event segmentation [66] show improved performance on machine learning benchmarks (see also [42,78,95]).

As with all cognitive modeling endeavors, further progress will depend on identifying specific sources of experimental data that can be used to constrain model development. In the Outstanding questions we list several domains that we think will be useful. One important source of constraints will be findings that shed light on interactions between EM and other memory systems. With regard to semantic memory, it will be useful to account for data on how prior knowledge benefits new learning [96,97], how offline replay of EMs updates semantic memory [98], and how semantic memory can blend together with episodic memory to yield constructive memory errors during retrieval ([99]; a recent model that addresses this last point is described in [40]); it will also be useful to account for how information is adaptively swapped between EM and working memory to support task performance [100–103]. In addition, MA-LLMs will likely benefit from incorporating the biological features of the hippocampus that support its ability to efficiently store and retrieve memories, many of which have been incorporated in other computational models [104–106]. Lastly, another frontier for constraining MA-LLMs is to consider input modalities other than text. Some newer models can process multimodal sensory input [107], and providing multimodal input to MA-LLMs [108] will extend the range of findings these models can address (e.g., relating to how visual changes can drive event segmentation and shape memory [58]). Going forward, we expect that appropriately leveraging these and other sources of data will further improve the fit between human EM and MA-LLMs, with strong benefits both for cognitive modeling and AI.

Acknowledgments

The authors gratefully acknowledge financial support from the Princeton Laboratory for Artificial Intelligence. The authors also thank Catherine Chen for her comments on an earlier draft of the manuscript and Hao Liu for helpful discussions.

Declaration of interests

No interests are declared.

References

1. Smith, T.A. *et al.* (2013) The context repetition effect: predicted events are remembered better, even when they don't happen. *J. Exp. Psychol. Gen.* 142, 1298–1308
2. Kim, G. *et al.* (2014) Pruning of memories by context-based prediction error. *Proc. Natl. Acad. Sci. U. S. A.* 111, 8997–9002
3. Kragel, J.E. and Voss, J.L. (2022) Looking for the neural basis of memory. *Trends Cogn. Sci.* 26, 53–65
4. Ryan, J.D. *et al.* (2000) Amnesia is a deficit in relational memory. *Psychol. Sci.* 11, 454–461
5. Ryan, J.D. and Cohen, N.J. (2004) The nature of change detection and online representations of scenes. *J. Exp. Psychol. Hum. Percept. Perform.* 30, 988–1015
6. Renoult, L. *et al.* (2019) From knowing to remembering: the semantic–episodic distinction. *Trends Cogn. Sci.* 23, 1041–1057
7. Lee, H. *et al.* (2020) What can narratives tell us about the neural bases of human memory? *Curr. Opin. Behav. Sci.* 32, 111–119
8. Chen, J. *et al.* (2017) Shared memories reveal shared structure in neural activity across individuals. *Nat. Neurosci.* 20, 115–125
9. Baldassano, C. *et al.* (2017) Discovering event structure in continuous narrative perception and memory. *Neuron* 95, 709–721
10. Barnett, A.J. *et al.* (2024) Hippocampal–cortical interactions during event boundaries support retention of complex narrative events. *Neuron* 112, 319–330

Outstanding questions

How well can MA-LLMs explain the influence of semantic memory on EM during encoding and retrieval? At encoding, new learning is improved by the presence of relevant prior knowledge; at retrieval, interactions between prior knowledge and EM have been shown to lead to errors in reconstructing past events. Any successful model of human memory will need to account for these positive and negative effects of prior knowledge on memory accuracy.

How can we better capture the influence of EM on semantic memory? In MA-LLMs, EM is often used to supplement the knowledge of a 'frozen', pretrained LLM, whereas – in the brain – EM is used to train semantic memory through a process of consolidation. Incorporating consolidation into MA-LLMs could help us to better understand how these systems mutually shape each other, and how the brain deals with potential redundancies, conflicts, or synergies between knowledge stored in these systems.

How well can MA-LLMs explain the dynamic interactions that occur between EM and working memory? Recent work has demonstrated that information is shuttled back and forth between an actively represented state (in working memory) and latent representation in EM depending on task demands. Data exploring these interactions can provide a rich source of constraints for bringing MA-LLMs closer to humans.

How can we leverage our knowledge of the biology of EM in humans to make MA-LLMs more human-like? Extant computational models of hippocampal contributions to EM posit a far more detailed set of pathways and mechanisms than the simple storage and retrieval systems used in MA-LLMs. Incorporating some of this hippocampal functionality into MA-LLMs could potentially improve both the fit to human data and model performance.

How do non-linguistic cues (e.g., changes in intonation) shape EM? This question can be addressed by incorporating multimodal LLMs that can handle visual and auditory input into MA-LLM architectures.

11. Ben-Yakov, A. and Henson, R.N. (2018) The hippocampal film editor: sensitivity and specificity to event boundaries in continuous experience. *J. Neurosci.* 38, 10057–10068
12. Chang, C.H.C. *et al.* (2021) Relating the past with the present: information integration and segregation during ongoing narrative processing. *J. Cogn. Neurosci.* 33, 1106–1128
13. Chen, J. *et al.* (2016) Accessing real-life episodic information from minutes versus hours earlier modulates hippocampal and high-order cortical dynamics. *Cereb. Cortex* 26, 3428–3441
14. Cohn-Sheehy, B.I. *et al.* (2021) The hippocampus constructs narrative memories across distant events. *Curr. Biol.* 31, 4935–4945
15. Michelmann, S. *et al.* (2021) Moment-by-moment tracking of naturalistic learning and its underlying hippocampo-cortical interactions. *Nat. Commun.* 12, 5394
16. Lu, Q. *et al.* (2022) A neural network model of when to retrieve and encode episodic memories. *eLife* 11, e74445
17. Giallanza, T. *et al.* (2024) Toward the emergence of intelligent control: episodic generalization and optimization. *Open Mind* 8, 688–722
18. Asai, A. *et al.* (2024) Reliable, adaptable, and attributable language models with retrieval. *arXiv*, Published online March 5, 2024. http://dx.doi.org/10.48550/arXiv.2403.03187
19. Wu, Y. *et al.* (2025) From human memory to AI memory: a survey on memory mechanisms in the era of LLMs. *arXiv*, Published online April 23, 2025. http://dx.doi.org/10.48550/arXiv.2504.15965
20. Raccah, O. *et al.* (2022) Memory in humans and deep language models: linking hypotheses for model augmentation. *arXiv*, Published online November 28, 2022. https://doi.org/10.48550/arXiv.2210.01869
21. Radford, A. *et al.* (2019) Language models are unsupervised multitask learners. *OpenAI Blog* 1, 9
22. Goldstein, A. *et al.* (2022) Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* 25, 369–380
23. Tuckute, G. *et al.* (2024) Language in brains, minds, and machines. *Annu. Rev. Neurosci.* 47, 277–301
24. Li, M. *et al.* (2024) NeedleBench: can LLMs do retrieval and reasoning in 1 million context window? *arXiv*, Published online July 16, 2024. http://dx.doi.org/10.48550/arXiv.2407.11963
25. Carlini, N. *et al.* (2023) Quantifying memorization across neural language models. In *Eleventh International Conference on Learning Representations*
26. Kandpal, N. *et al.* (2023) Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202) (Krause, A. *et al.*, eds), pp. 15696–15707, PMLR
27. Mallen, A. *et al.* (2023) When not to trust language models: investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Vol. 1) (Rogers, A. *et al.*, eds), pp. 9802–9822, Association for Computational Linguistics
28. Luo, Y. *et al.* (2024) An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv*, Published online April 2, 2024. http://dx.doi.org/10.48550/arXiv.2308.08747
29. McCloskey, M. and Cohen, N.J. (1989) Catastrophic interference in connectionist networks: the sequential learning problem. In *Psychology of Learning and Motivation* (Vol. 24) (Bower, G.H., ed.), pp. 109–165, Academic Press
30. Vaswani, A. *et al.* (2017) Attention is all you need. *Adv. Neural Inf. Proces. Syst.* 30, 5998–6008
31. Gu, A. and Dao, T. (2024) Mamba: linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*
32. Feng, L. *et al.* (2024) Were RNNs all we needed? *arXiv*, Published online November 28, 2024. http://dx.doi.org/10.48550/arXiv.2410.01201
33. Berglund, L. *et al.* (2024) The reversal curse: LLMs trained on 'A is B' fail to learn 'B is A'. In *Twelfth International Conference on Learning Representations*
34. Tay, Y. *et al.* (2022) Efficient transformers: a survey. *ACM Comput. Surv.* 55, 109
35. McClelland, J.L. *et al.* (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102, 419–457
36. Jiang, X. *et al.* (2024) Long term memory: the foundation of AI self-evolution. *arXiv*, Published online November 20, 2024. http://dx.doi.org/10.48550/arXiv.2410.15665
37. Gao, Y. *et al.* (2024) Retrieval-augmented generation for large language models: a survey. *arXiv*, Published online March 27, 2024. http://dx.doi.org/10.48550/arXiv.2312.10997
38. Xiao, C. *et al.* (2024) InfLLM: training free long-context extrapolation for LLMs with an efficient context memory. *Adv. Neural Inf. Proces. Syst.* 37, 119638–119661
39. Wu, Y. *et al.* (2022) Memorizing transformers. In *Tenth International Conference on Learning Representations*
40. Spens, E. and Burgess, N. (2024) Consolidation of sequential experience into a deep generative network explains human memory, prediction and planning. *bioRxiv*, Published online November 4, 2024. https://doi.org/10.1101/2024.11.04.621950
41. Barnes, C.A. *et al.* (1990) Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Prog. Brain Res.* 83, 287–300
42. Das, P. *et al.* (2024) Larimar: large language models with episodic memory control. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235) (Salakhutdinov, R. *et al.*, eds), pp. 10109–10126, PMLR
43. Li, B.Z. *et al.* (2025) Language modeling with editable external knowledge. In *Findings of the Association for Computational Linguistics (NAACL 2025)* (Chiruzzo, L. *et al.*, eds), pp. 3070–3090, ACL
44. Zhong, W. *et al.* (2024) MemoryBank: enhancing large language models with long-term memory. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence* (Vol. 38) (M. Wooldridge, et al., ed.), pp. 19724–19731, AAAI Press
45. Behrouz, A. *et al.* (2024) Titans: learning to memorize at test time. *arXiv*, Published online December 31, 2024. http://dx.doi.org/10.48550/arXiv.2501.00663
46. Munkhdalai, T. *et al.* (2024) Leave no context behind: efficient infinite context transformers with infini-attention. *arXiv*, Published online August 9, 2024. http://dx.doi.org/10.48550/arXiv.2404.07143
47. Spens, E. and Burgess, N. (2024) A generative model of memory construction and consolidation. *Nat. Hum. Behav.* 8, 526–543
48. Kumaran, D. *et al.* (2016) What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci.* 20, 512–534
49. Moscovitch, M. and Gilboa, A. (2024) Systems consolidation, transformation, and reorganization: multiple trace theory, trace transformation theory, and their competitors. In *The Oxford Handbook of Human Memory: Foundations and Applications* (Kahana, M.J. and Wagner, A.D., eds), pp. 1278–1328, Oxford University Press
50. van de Ven, G.M. *et al.* (2020) Brain-inspired replay for continual learning with artificial neural networks. *Nat. Commun.* 11, 4069
51. Mnih, V. *et al.* (2015) Human-level control through deep reinforcement learning. *Nature* 518, 529–533
52. Chukwu, E. and Bindschaedler, L. (2025) May the memory be with you: efficient and infinitely updatable state for large language models. In *Proceedings of the 5th Workshop on Machine Learning and Systems*, pp. 200–207, ACM
53. de Masson d'Autume, C. *et al.* (2019) Episodic memory in lifelong language learning. *Adv. Neural Inf. Proces. Syst.* 32, 13132–13141
54. Zacks, J.M. *et al.* (2007) Event perception: a mind/brain perspective. *Psychol. Bull.* 133, 273–293
55. Geerligs, L. *et al.* (2022) A partially nested cortical hierarchy of neural states underlies event segmentation in the human brain. *eLife* 11, e77430
56. Franklin, N.T. *et al.* (2020) Structured event memory: a neuro-symbolic model of event cognition. *Psychol. Rev.* 127, 327–361
57. Shin, Y.S. and DuBrow, S. (2021) Structuring memory through inference-based event segmentation. *Top. Cogn. Sci.* 13, 106–127
58. Fountas, Z. *et al.* (2022) A predictive processing model of episodic memory and time perception. *Neural Comput.* 34, 1501–1544

59. Kumar, M. *et al.* (2023) Bayesian surprise predicts human event segmentation in story listening. *Cogn. Sci.* 47, e13343
60. Reynolds, J.R. *et al.* (2007) A computational model of event segmentation from perceptual prediction. *Cogn. Sci.* 31, 613–643
61. Michelmann, S. *et al.* (2023) Evidence that event boundaries are access points for memory retrieval. *Psychol. Sci.* 34, 326–344
62. Clewett, D. *et al.* (2019) Transcending time in the brain: how event memories are constructed from experience. *Hippocampus* 29, 162–183
63. Izacard, G. *et al.* (2023) Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.* 24, 1–43
64. Lewis, P. *et al.* (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Proces. Syst.* 33, 9459–9474
65. Borgeaud, S. *et al.* (2022) Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning* (Vol. 162) (Chaudhuri, K. *et al.*, eds), pp. 2206–2240, PMLR
66. Fountas, Z. *et al.* (2025) Human-inspired episodic memory for infinite context LLMs. In *Thirteenth International Conference on Learning Representations*
67. Hahamy, A. *et al.* (2023) The human brain reactivates context-specific past information at event boundaries of naturalistic experiences. *Nat. Neurosci.* 26, 1080–1089
68. Huff, M. *et al.* (2014) Changes in situation models modulate processes of event perception in audiovisual narratives. *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 1377–1388
69. Zacks, J.M. *et al.* (2011) Prediction error associated with the perceptual segmentation of naturalistic events. *J. Cogn. Neurosci.* 23, 4057–4066
70. Ben-Yakov, A. *et al.* (2013) Hippocampal immediate poststimulus activity in the encoding of consecutive naturalistic episodes. *J. Exp. Psychol. Gen.* 142, 1255–1263
71. Ben-Yakov, A. and Dudai, Y. (2011) Constructing realistic engrams: poststimulus activity of hippocampus and dorsal striatum predicts subsequent episodic memory. *J. Neurosci.* 31, 9032–9042
72. Ram, O. *et al.* (2023) In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguist.* 11, 1316–1331
73. Jiang, Z. *et al.* (2023) Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Bouamor, H. *et al.*, eds), pp. 7969–7992, ACL
74. Healey, M.K. *et al.* (2019) Contiguity in episodic memory. *Psychon. Bull. Rev.* 26, 699–720
75. Howard, M.W. *et al.* (2015) A distributed representation of internal time. *Psychol. Rev.* 122, 24–53
76. Howard, M.W. and Kahana, M.J. (2002) A distributed representation of temporal context. *J. Math. Psychol.* 46, 269–299
77. Polyn, S.M. *et al.* (2009) A context maintenance and retrieval model of organizational processes in free recall. *Psychol. Rev.* 116, 129–156
78. Park, S. and Bak, J. (2024) Memoria: resolving fateful forgetting problem through human-inspired memory architecture. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235) (Salakhutdinov, R. *et al.*, eds), pp. 39587–39615, PMLR
79. Ji-An, L. *et al.* (2024) Linking in-context learning in transformers to human episodic memory. *Adv. Neural Inf. Proces. Syst.* 37, 6180–6212
80. Watkins, O.C. and Watkins, M.J. (1975) Buildup of proactive inhibition as a cue-overload effect. *J. Exp. Psychol. [Hum. Learn.]* 1, 442–452
81. Anderson, J.R. (1974) Retrieval of propositional information from long-term memory. *Cognit. Psychol.* 6, 451–474
82. Schechtman, E. *et al.* (2021) Multiple memories can be simultaneously reactivated during sleep as effectively as a single memory. *Commun. Biol.* 4, 25
83. Tal, A. *et al.* (2024) The reach of reactivation: effects of consciously triggered versus unconsciously triggered reactivation of associative memory. *Proc. Natl. Acad. Sci.* 121, e2313604121
84. Asai, A. *et al.* (2024) Self-RAG: learning to retrieve, generate, and critique through self-reflection. In *Twelfth International Conference on Learning Representations*
85. Glass, M. *et al.* (2022) Re2G: retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Carpuat, M. *et al.*, eds), pp. 2701–2715, ACL
86. Yoran, O. *et al.* (2024) Making retrieval-augmented language models robust to irrelevant context. In *Twelfth International Conference on Learning Representations*
87. Wu, S. *et al.* (2024) How easily do irrelevant inputs skew the responses of large language models? In *First Conference on Language Modeling*
88. Kwiatkowski, T. *et al.* (2019) Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguist.* 7, 453–466
89. Lohnas, L.J. *et al.* (2015) Expanding the scope of memory search: modeling intralist and interlist effects in free recall. *Psychol. Rev.* 122, 337–363
90. Pink, M. *et al.* (2024) Assessing episodic memory in LLMs with sequence order recall tasks. *arXiv*, Published online October 10, 2024. http://dx.doi.org/10.48550/arXiv.2410.08133
91. Huet, A. *et al.* (2025) Episodic memories generation and evaluation benchmark for large language models. In *Thirteenth International Conference on Learning Representations*
92. Yao, S. *et al.* (2022) WebShop: towards scalable real-world web interaction with grounded language agents. *Adv. Neural Inf. Proces. Syst.* 35, 20744–20757
93. Maharana, A. *et al.* (2024) Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1) (Ku, L.W. *et al.*, eds), pp. 13851–13870, ACL
94. Zhang, Z. *et al.* (2024) A survey on the memory mechanism of large language model based agents. *arXiv*, Published online April 21, 2024. http://dx.doi.org/10.48550/arXiv.2404.13501
95. Jiménez Gutiérrez, B. *et al.* (2024) HippoRAG: neurobiologically inspired long-term memory for large language models. *Adv. Neural Inf. Proces. Syst.* 37, 59532–59569
96. Guerreiro, I.C. and Clopath, C. (2024) Memory's gatekeeper: the role of PFC in the encoding of congruent events. *Proc. Natl. Acad. Sci.* 121, e2403648121
97. McClelland, J.L. (2013) Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *J. Exp. Psychol. Gen.* 142, 1190–1210
98. Singh, D. *et al.* (2022) A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation. *Proc. Natl. Acad. Sci.* 119, e2123432119
99. Schacter, D.L. *et al.* (1998) The cognitive neuroscience of constructive memory. *Annu. Rev. Psychol.* 49, 289–318
100. Beukers, A.O. *et al.* (2021) Is activity silent working memory simply episodic memory? *Trends Cogn. Sci.* 25, 284–293
101. Hasson, U. *et al.* (2015) Hierarchical process memory: memory as an integral component of information processing. *Trends Cogn. Sci.* 19, 304–313
102. Honey, C.J. *et al.* (2023) Psychological momentum. *Curr. Dir. Psychol. Sci.* 32, 284–292
103. Momennejad, I. *et al.* (2021) Rational use of episodic and working memory: a normative account of prospective memory. *Neuropsychologia* 158, 107657
104. Chandra, S. *et al.* (2025) Episodic and associative memory from spatial scaffolds in the hippocampus. *Nature* 638, 739–751
105. Schapiro, A.C. *et al.* (2017) Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. B Biol. Sci.* 372, 20160049
106. Whittington, J.C.R. *et al.* (2020) The Tolman–Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell* 183, 1249–1263.e23
107. Zhang, D. *et al.* (2024) MM-LLMs: recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics (ACL 2024)*, pp. 12401–12430
108. Lin, Y. *et al.* (2025) HippoMM: hippocampal-inspired multimodal memory for long audiovisual event understanding. *arXiv*, Published online April 14, 2025. http://dx.doi.org/10.48550/arXiv.2504.10739

109. Bahdanau, D. *et al.* (2015) Neural machine translation by jointly learning to align and translate. In *Third International Conference on Learning Representations*
110. Hintzman, D.L. (1986) 'Schema abstraction' in a multiple-trace memory model. *Psychol. Rev.* 93, 411–428
111. Ramsauer, H. *et al.* (2021) Hopfield networks is all you need. In *Ninth International Conference on Learning Representations*
112. Gershman, S.J. *et al.* (2025) Key–value memory in the brain. *Neuron* 113, 1694–1707.e1
113. Whittington, J.C.R. *et al.* (2022) Relating transformers to models and neural representations of the hippocampal formation. In *Tenth International Conference on Learning Representations*
114. Zuo, X. *et al.* (2020) Temporal integration of narrative information in a hippocampal amnesic patient. *NeuroImage* 213, 116658
115. Robertson, S. and Zaragoza, H. (2009) The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.* 3, 333–389
116. Izacard, G. *et al.* (2022) Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*
117. Ritvo, V.J.H. *et al.* (2019) Nonmonotonic plasticity: how memory retrieval drives learning. *Trends Cogn. Sci.* 23, 726–742
118. Lee, J.L.C. *et al.* (2017) An update on memory reconsolidation updating. *Trends Cogn. Sci.* 21, 531–545
119. Gershman, S.J. *et al.* (2017) The computational nature of memory modification. *eLife* 6, e23763
120. Duncan, K.D. and Schlichting, M.L. (2018) Hippocampal representations as a function of time, subregion, and brain state. *Neurobiol. Learn. Mem.* 153, 40–56
121. Wanjia, G. *et al.* (2021) Abrupt hippocampal remapping signals resolution of memory interference. *Nat. Commun.* 12, 4816
122. Ritvo, V.J. *et al.* (2024) A neural network model of differentiation and integration of competing memories. *eLife* 12, RP88608
123. Zadbood, A. *et al.* (2022) Neural representations of naturalistic events are updated as our understanding of the past changes. *eLife* 11, e79045
124. Zacks, J.M. (2020) Event perception and memory. *Annu. Rev. Psychol.* 71, 165–191
125. Anderson, M.C. (2003) Rethinking interference theory: executive control and the mechanisms of forgetting. *J. Mem. Lang.* 49, 415–445
126. Schrimpf, M. *et al.* (2021) The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci.* 118, e2105646118
127. Toneva, M. and Wehbe, L. (2019) Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Adv. Neural Inf. Proces. Syst.* 32, 14954–14964
128. Caucheteux, C. *et al.* (2023) Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat. Hum. Behav.* 7, 430–441
129. Jain, S. and Huth, A. (2018) Incorporating context into language encoding models for fMRI. *Adv. Neural Inf. Proces. Syst.* 31, 6628–6637
130. Tikochinski, R. *et al.* (2025) Incremental accumulation of linguistic context in artificial and biological neural networks. *Nat. Commun.* 16, 803