Evaluating Models of Naturalistic Episodic Memory

Mathis Pink (mpink@mpi-sws.org)
Max Planck Institute for Software Systems
Saarbrücken, Germany

Shashwat Saxena (ssaxena@mpi-sws.org)
Max Planck Institute for Software Systems
Saarbrücken, Germany

Mariya Toneva (mtoneva@mpi-sws.org)
Max Planck Institute for Software Systems
Saarbrücken, Germany

Abstract

Recent advances in large language models (LLMs) have enabled them to process extended naturalistic inputs, making them promising candidates for modeling human episodic memory. However, standard transformer-based LLMs rely on full self-attention and positional encoding, which diverge from human episodic memory by supporting soft, parallel attention over complete input sequences. EM-LLM, a recent modification, inspired by episodic memory, replaces full attention with episodic retrieval from a non-parametric memory filled with discrete past episodes that were segmented via surprisal. Here, we evaluate whether EM-LLM captures a core property of episodic memory: the ability to recall the temporal order of events. Using a recency judgment task on segments from a full-length novel and comparing to human behavioral data, we find that a standard full-attention LLM aligns with human performance, while EM-LLM fails to recover temporal order across long sequences. These findings reveal a key limitation in EM-LLM's current design and suggest that temporal organization may require either additional architectural biases or learned representations—highlighting new directions for modeling episodic memory in naturalistic contexts.

Keywords: Episodic Memory, Computational Modeling, LLMs, Temporal Order Memory, Recency Judgements, Narrative Memory

Introduction

Theory-driven computational models of episodic memory, such as the Temporal Context Model (TCM) (Howard & Kahana, 2002), the Context Maintenance and Retrieval (CMR) model (Polyn et al., 2009), or the Tolman-Eichenbaum Machine (Whittington et al., 2020) allow the modeling of episodic memory in simplified, laboratory-based settings. This limits their ability to capture the semantic richness and contextual complexity of real-world episodic experiences. Large language models (LLMs) offer a potential data-driven tool to overcome these shortcomings, enabling the processing of semantically rich, extended sequences, such as narratives in full-length books. Indeed, attention mechanisms in LLMs have recently been linked to models of episodic memory (Ji-An et al., 2024), highlighting that these models may present viable directions for future modeling.

However, the data-driven nature of these models does not come with inductive biases that mirror what is known about episodic memory. This leads to important differences between LLM's main computational mechanism–full self-attention–and human episodic memory, as it involves simultaneously attending over the entire past rather than selectively retrieving specific episodes. The lack of strong inductive biases for discrete retrieval highlights that a core aspect of episodic memory in humans and other animals is missing in current LLMs.

A recently proposed replacement for the full self-attention in transformers aims to address these shortcomings by taking algorithmic inspiration from episodic memory (Fountas et al., 2025). The method, called EM-LLM, maintains a non-parametric memory that is iteratively filled with episode representations from a local context window, segmented based on prediction errors. In every attention layer, this replaces full attention over the past with attention over discretely retrieved episodes to contextualize a short local context window. EM-LLM is built on top of an existing LLM and does not require any re-training of the underlying language model.

In this paper, we evaluate whether EM-LLM already presents a promising tool for modeling human episodic memory for natural stimuli. By investigating the behavioral alignment with human performance patterns on a recency judgment task performed over segments taken from a complete book of close to 70k words (Pink et al., 2024), we show that EM-LLM fails to recall the temporal order of narrative segments, while full attention mirrors human performance. We suggest directions to improve this method as a tool for the modeling of human episodic memory, and propose to use tools from mechanistic interpretability to gain mechanistic insights from the full attention model.

Methods

Recency Judgment Task and Human Data

Task. Recency judgment tasks provide a method to evaluate episodic memory by requiring experiment participants to recall the order of two items in a previously seen sequence (Eichenbaum, 2013; Davachi & DuBrow, 2015). Here, we adopt this paradigm but focus on modeling recency judgments of segments taken from a full-length book.

Human experiment data. We compare models to behavioral data from Pink et al. (2024), who tested participants after they had read the novel "The Murder of Roger Ackroyd" by Agatha Christie within the past 30 days. We focus on a subset of this dataset here, which includes 296 pairs of 50-word segments extracted from the book (excluding chapter titles). The distance between the segment pairs in the book ranges from 12 words to around 34,000 words, and in total, the data comprises 990 recency judgments on which segment in each pair appeared earlier in the book. These judgments were provided by 97 participants.

Models and Attention Mechanisms

We evaluate Llama3.1-8b-Instruct in two settings on the same task and data as our human long-term memory experiment: standard full self-attention and with EM-LLM replacing attention layers. We adopt the prompting methodology of Pink et al. (2024), who proposed Sequence Order Recall Task (SORT), a recency judgment benchmark for testing long-term memory in LLMs. For each of the 296 unique segment pairs, the models are prompted twice: once with the segments in the correct chronological order, and once in reverse.

Llama3.1-8b-Instruct. We evaluate Llama3.1-8b-Instruct (Grattafiori et al., 2024), an LLM that supports attention over 128k tokens in its context window. As many other recent

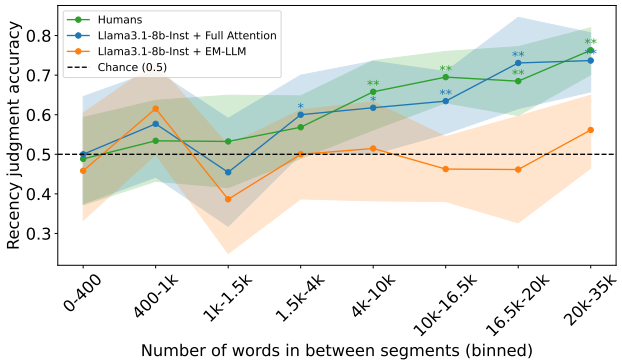

Figure 1: Recency judgment accuracy with increasing distances between pairs of segments. Llama3.1-8b-Instruct with full attention closely aligns with human performance, while the same model with EM-LLM does not perform above chance level. Asterisks indicate conventional levels of statistical significance, and the shaded areas represent 95% confidence intervals.

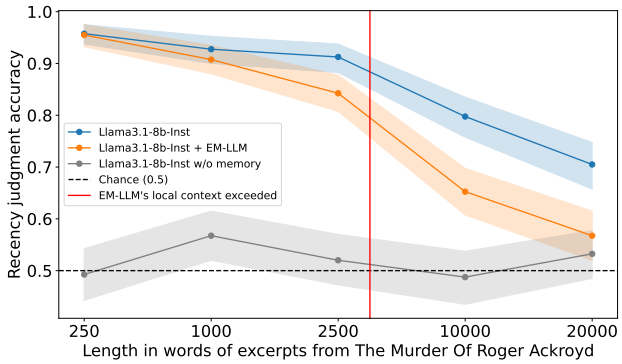

Figure 2: Recency judgment accuracy for inputs with increasing lengths. The red line indicates when EM-LLM has to rely on retrieval of episodes rather than its local context window.

LLMs, it uses Rotary Positional Encodings (RoPE) in each attention layer (Su et al., 2023). Unlike in previous models like GPT-2 (Radford et al., 2019), RoPE does not add positional vectors to values, but instead rotates query and key representations. Thereby positional encodings can only affect attention scores but can not be directly attended themselves (and thus can also not trivially be copied for recency judgments).

EM-LLM as a replacement to full self-attention. EM-LLM (Fountas et al., 2025) replaces standard attention computation with discrete retrieval of contiguous episode representations—*without any fine-tuning of the original model*. Instead of attending to all previous tokens in parallel, the model processes input sequentially using a small local context window. At each step, each attention layer retrieves a set of previously segmented episodes, based on surprisal-driven boundaries. Retrieval is performed by comparing representative tokens from each episode to those in the current context window, using a similarity-based approximate nearest-neighbor search.

Since the local context window is relatively small (4096 tokens in our experiments), EM-LLM processes long sequences—such as the full novel containing 94k tokens—by iteratively sliding this window over the entire input. Within the local window, episodes are segmented at tokens where a prediction error exceeds a threshold (set to 1 in our experiments). Each attention layer retrieves a set of past episodes—up to 4096 tokens in total.

Results

Full attention closely matches human performance on recency judgments across a full book. We find that Llama3.1-8b-Instruct, when equipped with full causal self-attention, performs comparably to human participants who read the book within the past 30 days (see Fig. 1). This aligns with prior work linking transformer attention mechanisms to human episodic memory processes (Ji-An et al., 2024; Whittington et al., 2022, 2025; Ellwood, 2024; Park et al., 2025).

EM-LLM performs at chance on recency judgments when evaluated over the full book. In contrast to full-attention models, EM-LLM is unable to recall the order of segments when the entire book is included, performing no better than chance (see Fig. 1). This pattern persists even when the relevant episodes are successfully retrieved.

EM-LLM can perform recency judgments when relevant content is within its local context window. Using a subset of the BookSORT dataset from Pink et al. (2024) (filtered for The Murder of Roger Ackroyd, and 50-word segments), we find that EM-LLM is generally capable of recency judgments, with performance comparable to full-attention models, *but only when* a sufficiently large portion of the relevant book excerpt remains within the local context window (see Fig. 2).

Discussion

Although EM-LLM can retrieve relevant representations from long contexts, it fails to support recency judgments—likely due to the absence of temporal information in the retrieved episode representations. Full-attention models retain a global positional code to guide attention scores (Su et al., 2023), thereby possibly embedding information similar to that in time cells in the hippocampus (Salz et al., 2016), which could be what enables temporal order memory in our experiment. In contrast, EM-LLM appears to lack an effective temporal scaffold beyond its local context window. Future work could seek to improve EM-LLM as a model of episodic memory by embedding additional temporal structure via explicitly given relative positional encodings or enabling a learnable emergent temporal organization. Meanwhile, the striking behavioral similarity between causal attention and episodic memory for recency judgment tasks can be explained with methods from mechanistic interpretability.

Acknowledgements

Funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 “Neuroexplicit Models of Language, Vision, and Action” - project number 471607914.

References

Davachi, L., & DuBrow, S. (2015). How the hippocampus preserves order: the role of prediction and context. *Trends in cognitive sciences*, *19*(2), 92–99.

Eichenbaum, H. (2013). Memory on time. *Trends in cognitive sciences*, *17*(2), 81–88.

Ellwood, I. T. (2024, January). Short-term hebbian learning can implement transformer-like attention. *PLOS Computational Biology*, *20*(1), e1011843. Retrieved from http://dx.doi.org/10.1371/journal.pcbi.1011843 doi: 10.1371/journal.pcbi.1011843

Fountas, Z., Benfeghoul, M., Oomerjee, A., Christopoulou, F., Lampouras, G., Ammar, H. B., & Wang, J. (2025). Human-inspired episodic memory for infinite context LLMs. In *The thirteenth international conference on learning representations.* Retrieved from https://openreview.net/forum?id=BI2int5SAC

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... Ma, Z. (2024). *The llama 3 herd of models.* Retrieved from https://arxiv.org/abs/2407.21783

Howard, M. W., & Kahana, M. J. (2002, June). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299. Retrieved from http://dx.doi.org/10.1006/jmps.2001.1388 doi: 10.1006/jmps.2001.1388

Ji-An, L., Zhou, C. Y., Benna, M. K., & Mattar, M. G. (2024). Linking in-context learning in transformers to human episodic memory. In *The thirty-eighth annual conference on neural information processing systems.* Retrieved from https://openreview.net/forum?id=AYDBFxNon4

Park, C. F., Lee, A., Lubana, E. S., Yang, Y., Okawa, M., Nishi, K., ... Tanaka, H. (2025). ICLR: In-context learning of representations. In *The thirteenth international conference on learning representations.* Retrieved from https://openreview.net/forum?id=pXlmOmlHJZ

Pink, M., Vo, V. A., Wu, Q., Mu, J., Turek, J. S., Hasson, U., ... Toneva, M. (2024). *Assessing episodic memory in llms with sequence order recall tasks.* Retrieved from https://arxiv.org/abs/2410.08133

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156. Retrieved from http://dx.doi.org/10.1037/a0014420 doi: 10.1037/a0014420

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Salz, D. M., Tiganj, Z., Khasnabish, S., Kohley, A., Sheehan, D., Howard, M. W., & Eichenbaum, H. (2016, July). Time cells in hippocampal area ca3. *Journal of Neuroscience*, *36*(28), 7476–7484. Retrieved from http://dx.doi.org/10.1523/JNEUROSCI.0087-16.2016 doi: 10.1523/jneurosci.0087-16.2016

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2023). *Roformer: Enhanced transformer with rotary position embedding.* Retrieved from https://arxiv.org/abs/2104.09864

Whittington, J. C. R., Dorrell, W., Behrens, T. E. J., Ganguli, S., & El-Gaby, M. (2025, January). A tale of two algorithms: Structured slots explain prefrontal sequence memory and are unified with hippocampal cognitive maps. *Neuron*, *113*(2), 321–333.e6. Retrieved from http://dx.doi.org/10.1016/j.neuron.2024.10.017 doi: 10.1016/j.neuron.2024.10.017

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020, November). The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, *183*(5), 1249–1263.e23. Retrieved from http://dx.doi.org/10.1016/j.cell.2020.10.024 doi: 10.1016/j.cell.2020.10.024

Whittington, J. C. R., Warren, J., & Behrens, T. E. (2022). Relating transformers to models and neural representations of the hippocampal formation. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=B8DVo9B1YE0