A Unified Benchmark for Human-Like Memory in Artificial Agents

Lucas Gruaz (lucas.gruaz@epfl.ch)
Brain-Mind Institute, School of Life Sciences
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland

Aude Maier (aude.maier@epfl.ch)
Brain-Mind Institute, School of Life Sciences
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland

Johanni Brea (johanni.brea@epfl.ch)
Brain-Mind Institute, School of Life Sciences
School of Computer and Communication Sciences
EPFL, Lausanne, Switzerland

Abstract

Human memory exhibits a diverse range of well-documented phenomena, including forgetting curves, interference effects, and schema-based distortions. While existing computational models attempt to capture aspects of these phenomena, they are often evaluated in isolation using task-specific experimental setups, limiting their generalizability and comparability.

We develop a unified benchmark for systematically evaluating memory models based on their ability to reproduce human-like memory phenomena. Our approach includes: (1) analyzing and formalizing a diverse set of memory phenomena in generalizable terms, independent of specific experimental paradigms, and (2) developing an evaluation framework that tests these phenomena within a common environment. This allows to test all phenomena on the same memory-augmented agent.

We test different memory models on schema-based distortion, memory conjunction errors, repetition and serial position effects. We find that none of the tested memory models qualitatively matches human memory behavior on all these phenomena, and we identify promising directions for future research on memory models.

Keywords: episodic memory; benchmark; schema-based distortion; memory conjunction errors; contiguity; recency; LSTM; CMR; VAE

Introduction

Human memory is influenced by various factors that affect how experiences are encoded, stored, and retrieved. These include contextual cues, prior knowledge, and the limitations of storage and retrieval processes. These factors give rise to well-documented memory phenomena, such as forgetting effects (Kahana, Diamond, & Aka, 2022), structural biases (e.g., schema-based distortions (Carmichael, Hogan, & Walter, 1932; Alba & Hasher, 1983; van Kesteren, Ruiter, Fernández, & Henson, 2012)), associative errors (e.g., memory conjunction errors (Kroll, Knight, Metcalfe, Wolf, & Tulving, 1996; Rubin, Petten, Glisky, & Newberg, 1999)), and temporal dynamics (e.g., recency (Murdock Jr., 1962) and contiguity effects (Kahana et al., 2022)). Understanding these effects is essential for developing computational models that accurately reflect human memory.

Despite extensive research in psychology and neuroscience, computational memory models often focus on a narrow subset of these phenomena, leading to specialized models that fail to generalize across different tasks. To address this, we introduce a benchmark designed to systematically evaluate memory phenomena in artificial agents using a common framework for testing across multiple tasks.

Methods

To evaluate memory phenomena in artificial agents, we design a framework in which a model observes sequences of events and produces outputs to solve a variety of tasks requiring memory. Although text- or image-based representations of events seem like natural choices, both present significant limitations. Text's discrete nature prevents testing certain memory phenomena involving distortion, while image-based environments involve complex perceptual processing that can obscure the core memory mechanisms under investigation. Both modalities are also constrained by high computational costs.

To address these limitations, we use generated event sequences embedded in a continuous vector space. This approach allows for controlled variability and precise manipulation of event similarity and structure, enabling systematic evaluation of memory phenomena. Our framework is computationally efficient, flexible, and scalable. It supports both supervised and reinforcement learning, and it enables simple sequence parsing for non-trainable models.

Event generation

Events are generated from a vocabulary that stores schemas and their representations in $[0,1]^n$, as depicted in Figure 1A. When generating an event, a schema is selected, either randomly or following some sequence structure, and Gaussian

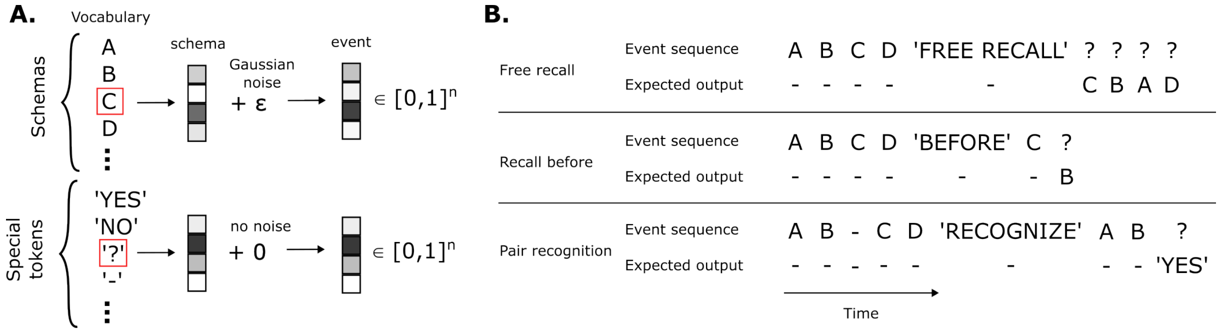

Figure 1: Event representation and task design. **A.** A vocabulary stores a list of schemas and their vector representation. When generating an event, a schema is selected and Gaussian noise is added. Special tokens include responses ('YES', 'NO', '-') and task instructions ('FREE RECALL', 'BEFORE', 'RECOGNIZE', '?'). **B.** During training, sequences of events are shown to the agent, followed by task instructions and output prompts. The expected outputs depend on the task and the event sequence. Any permutation of 'A', 'B', 'C', 'D' counts as correct in 'FREE RECALL' training.

noise is added to introduce variability. The vocabulary also includes special tokens for task instructions, ensuring consistency across different memory tasks.

Training tasks

Training tasks consist of sequences of events, followed by task instructions. All inputs and outputs are expressed in $[0,1]^n$. The benchmark includes various tasks designed to develop the memory agent's capabilities without explicitly training on the experiments used to assess memory phenomena. Examples of training tasks are presented in Figure 1B.

Phenomena testing

The benchmark assesses models based on their ability to qualitatively replicate human-like memory phenomena. Testing tasks follow the same format as training tasks but are specifically designed to assess particular memory phenomena. We currently focus on six key effects:

- **schema-based distortion:** recalled events shift toward high-density regions of the event space. We assess this in 'Recall before' test sequences.
- **memory conjunction errors:** false recognition of events that combine features of previously observed events. We assess this with 'Pair recognition' test sequences.
- **recency:** in free recall and recognition, recent events are more likely to be correctly recalled.
- **primacy:** in free recall, events in the beginning of the sequence are more likely to be recalled.
- **contiguity:** in free recall, temporally adjacent events tend to be recalled together.
- **repetition:** in free recall, repeated events are more likely to be recalled.

Each model is evaluated in multiple trials, and the performance is measured using statistical tests that determine how significantly each effect appears above the chance level.

Results

So far, we evaluated existing models on our benchmark, including a temporal context model (CMR, (Polyn, Norman, & Kahana, 2009), using the distributed implementation available here), a transformer (Vaswani et al., 2017), recurrent neural networks (LSTM, (Hochreiter & Schmidhuber, 1997), xLSTM (Beck et al., 2024) Mamba (Gu & Dao, 2024), DNC (Graves, Wayne, Danihelka, et al., 2016)), and a variational autoencoder (VAE)-based model (Kingma & Welling, 2013), inspired by recent work on generative human memory modeling (Nagy, Török, & Orbán, 2020; Fayyaz et al., 2022; Spens & Burgess, 2024). VAEs focus primarily on encoding and decoding, lacking the capacity to solve the complex tasks in our benchmark. To address this, we paired the VAE with a simple oracle agent that always produces the correct output, enabling us to assess whether the VAE component captures relevant memory phenomena.

Table 1 summarizes the statistical significance of each model's ability to replicate key memory phenomena described in Phenomena testing. The CMR model, designed for free recall tasks, is limited to that context and cannot generalize to other settings. Recurrent neural networks are more flexible but only exhibit a subset of the memory phenomena. The VAE captures schema-based distortion but cannot be directly applied to all tasks. Overall, while some models captured specific effects, none fully reproduced human memory across all tested phenomena.

Conclusion

Our unified benchmark systematically tests memory effects across different tasks, providing a framework for assessing computational memory models. So far, we tested three representative types of memory models from different fields of memory research (recurrent neural networks: machine learning, CMR: computational model of classical lab experiments, VAE: computational model of generative memory), and showed that none of these models can replicate all the tested

Table 1: Statistical evidence for human-like memory phenomena in artificial agents. For each model, we report the proportion of 100-sample simulations in which effects are statistically significant ($p < 0.01$). The symbol ”na” indicates that the model cannot be trained or evaluated on the corresponding task.

	CMR	VAE-based	Transformer	LSTM	xLSTM	Mamba	DNC
SB distortion	nan	**1.0**	**1.0**	0.02	0.02	0.01	**0.75**
Conjunction errors	nan	0.0	**1.0**	**0.85**	**0.86**	**0.8**	**0.74**
Recency (recognition)	nan	0.0	0.01	0.0	0.0	0.07	0.0
Recency (free recall)	**1.0**	0.0	0.0	0.0	0.0	**1.0**	0.46
Primacy	**0.81**	0.0	0.0	**1.0**	**1.0**	0.0	**1.0**
Contiguity	**1.0**	0.01	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
Repetition	0.12	0.0	**1.0**	**1.0**	**1.0**	**0.79**	**1.0**

memory phenomena, indicating that we are lacking a unified model of human memory. During the upcoming months until the CCN conference, we will expand the benchmark to include additional memory effects, evaluate more sophisticated models, and introduce more training tasks to develop generalized memory representations.

Acknowledgments

This work was supported by the Swiss National Science Foundation grants 200020-207426 and 200021-236436.

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*, 203–231.

Beck, M., et al. (2024, May). Xlstm: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*.

Carmichael, L., Hogan, H. P., & Walter, A. A. (1932). An experimental study of the effect of language on the reproduction of visually perceived form. *Journal of Experimental Psychology*, *15*, 73–86.

Fayyaz, Z., Altamimi, A., Zoellner, C., Klein, N., Wolf, O. T., Cheng, S., & Wiskott, L. (2022). A Model of Semantic Completion in Generative Episodic Memory. *Neural Computation*, *34*, 1841–1870.

Graves, A., Wayne, G., Danihelka, I., et al. (2016, Oct). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471–476. doi: 10.1038/nature20101

Gu, A., & Dao, T. (2024, May). *Mamba: Linear-time sequence modeling with selective state spaces.* (Preprint)

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*, 1735–1780.

Kahana, M. J., Diamond, N. B., & Aka, A. (2022). Laws of Human Memory. *PsyArXiv Preprints*.

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes.

Kroll, N. E., Knight, R. T., Metcalfe, J., Wolf, E. S., & Tulving, E. (1996). Cohesion Failure as a Source of Memory Illusions. *Journal of Memory and Language*, *35*, 176–196.

Murdock Jr., B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488.

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

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*, 129–156.

Rubin, S. R., Petten, C. V., Glisky, E. L., & Newberg, W. M. (1999). Memory conjunction errors in younger and older adults: Event-related potential and neuropsychological data. *Cognitive Neuropsychology*, *16*, 459–488.

Spens, E., & Burgess, N. (2024). A generative model of memory construction and consolidation. *Nature Human Behaviour*, *8*, 526–543.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems (neurips).*