nature portfolio

Peer Review File

A neural network model of free recall learns multiple memory strategies

Corresponding Author: Professor Marcelo Mattar

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors present a simulation analysis of a set of bespoke computational models trained to optimize performance in free recall tasks. They find that, in addition to the well-documented Temporal Context strategy, characterized by relatively symmetric conditional recall probabilities, some of the networks, under certain conditions, develop what they term a "memory palace"-like strategy.

Overall the findings are very interesting, and generate many testable predictions for real-world experiments on the formation of memory palaces, while also proposing constraints for the utility of this widely celebrated strategy. Our comments focus mainly on tamping down or redirecting the interpretations and claims, and strengthening the connection to both existing TCM theory and empirically observed phenomena.

1. The main concern I have is that the way temporal context is operationalized here leaves out something important in how it has been conceived in previous studies. Drifting temporal context doesn't just emerge entirely from internal states – it also consists of representations of (temporally evolving) external stimuli (e.g. physical location, temporal/rhythmic fluctuations in light and background noise). What would happen if you include this sort of information in the previously one-hot encoded input (usually this is done as a vector concatenated to the stimulus representation)?

2. The fact that flushing WM leads to memory palace – suggests that to maximize the efficacy of this strategy people should study in different environments than they retrieve, a counterintuitive but exciting finding. Is this finding robust to the inclusion of a more traditional context vector?

3. Were other naturalistic strategies possible given these impoverished stimulus classes? e.g. clustering by conceptual/featural similarity? (I remember heart suit cards mostly together, etc) One way to examine this may be to ask whether similar stimuli (e.g. by cosine distance) are retrieved together. If not, is there a reason that these strategies never developed? Might they have developed with more rich stimuli that had been processed more categorically by something like the visual processing hierarchy in brains?

4. Fig 1G seems to operationalize the distinction between temporal context and memory palace in that the latter should show no backward recall. Is this strict forward processing actually observed in real memory palace-driven behavior? Do individuals not sometimes "retrace their steps?" From a quick search there appears not to be much laboratory work on this, but videos of memory champions suggest that it happens at least sometimes – indeed it appears to be encouraged by some practitioners as a way of building a robust "palace."

5. Relatedly, do humans ever exhibit such extreme forward-only recall in free recall tasks, absent the explicit instruction to start from the beginning, given ahead of study?

6. It is commonly thought that memory palaces are supported by spatial maps. Did the authors observe evidence in the internal activations of the trained networks that are consistent with map-like representations? Alternatively, are the representations instead better thought of as directed line-graphs? Would more naturalistic representations yield memory palaces under different conditions?

7. Similarly, the idea of memory palaces usually involves scaffolding on _existing_, well learned, often personally meaningful spatial arrangements, not just a simplistic ordinal position. Can the authors discuss some of the potential differences between their ordinal code and the one they observed here? Propose an architecture that would be able to leverage this kind of knowledge?

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
Li and colleagues present a creative study of human memory search that bridges with the machine intelligence literature in a novel and valuable way. In my mind, this paper makes a substantial contribution to both cognitive psychology theory and machine intelligence theory. In theories of human memory, the free-recall task provides valuable empirical data regarding the dynamic processes underlying memory search. One of the foundational computational models in this domain is the TCM model described in the introduction and used as a reference point throughout the paper. TCM is one of a class of retrieved-context models, and described many important behavioral phenomena, including the phenomenon at the center of this paper, regarding the temporal organization of human memory. In a sense, TCM is an example of a hand-crafted neural network model, designed to be simple, with very few free parameters. This makes the mechanisms of the model more interpretable, but also can limit the flexibility of the model. A major contribution of the current paper is showing that modern RNN technology can effectively learn to perform memory search in a variety of ways, reflecting the different ways human participants are known to approach memory search.

Making bridges between the modern technologies of machine intelligence, and the decades-long history of theoretical development of cognitive memory theory will have benefits for both domains. Here, the authors demonstrate that their novel modeling framework can learn to produce multiple kinds of internal representational structures and behaviors seen in humans. These span from serial recall with their connection to the memory palace technique, and to free recall, with 2 forms of behavior evident (forward biased and backward biased transitions).

The analysis of positional vs temporal internal states is excellent and has the potential to bridge two historically distinct lines of theory development in cognitive and mathematical psychology.

The addition of a memory module to the GRU system is innovative, in that a set of hidden states are formed during the simulated study period, much like the stored representations in a modern Hopfield model, or more aptly for the cognitive psychological memory literature, Hintzman's MINERVA model which precedes modern Hopfield by decades but has the same basic architecture and purpose.

The final section examining how an alternative recall task can alter the optimal strategy is also innovative and theoretically important. It is understandably shorter than the other sections and I look forward to seeing it expanded in future work.

Comments

Given the emphasis on human behavior, the paper would benefit from a figure showing examples of the different free recall strategies pulled from existing empirical data sets. Unless I'm missing something, all the strategy examples are pulled from the network behavior. Could have a separate figure demonstrating empirical data showing these things. I imagine the PEERS data set has all 3 in spades across participants. I don't think it would be necessary to integrate the empirical data into the existing analysis framework though, it seems cleaner as is, showing that these patterns arise naturally from the task/training demands.

Section 4.6 in the methods, I can't find mention to this in the main paper, though it seems a quite valuable description making an explicit connection between this work and modern transformer theory. I recommend at least adding some forward references to this section in the main body of the paper, if not moving some version of the section to earlier in the paper.

Minor comments

Would be interesting to hear a brief thought in conclusions about extending the model to situations where the model components are not wiped clean between trials.

Figure 2A caption. "Colors indicate strategy groups" -- clarify what each colored dot represents. "dots show reservoir RNN samples" -- does this mean black dots? There are colored and black dots, could clarify for the reader.

page 13. Temporal factor. Small correction: This technique was introduced by Polyn, Norman & Kahana Psych Review (2009) described on page 38. It was used the following year by Sederberg et al. (2010).

Temporal factor name. This analysis was initially called temporal factor, but in more recent years some groups have shifted to calling it a 'temporal organization score' or the more unwieldy 'percentile rank temporal organization score'. Temporal factor is more succinct, but there's a minor negative in that it can be misinterpreted as a form of factor analysis. Hong et al. (2023) JEP:LMC (The modulation and elimination of temporal org in free recall) provides a detailed description of the algorithm and relates it to the lag-CRP algorithm also used in this paper.

Signed by
Sean Polyn

(Remarks on code availability)
I have reviewed the code in a very limited sense, by clicking through the GitHub repository. I didn't run the code, but in looking through it, it is clearly well organized in a way that will support use by other researchers.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I thank the reviewers for their comprehensive response to my concerns.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors did an outstanding job responding to the reviews. I believe the manuscript is much strengthened as a result, and will make an excellent contribution to the literature.

The only minor thing I noted; with regard to the inclusion of the PEERS data in the extended figure identifying participants that indeed show an extreme version of the forward chaining / memory palace behavior in the lag-CRP. There is a paper by Romani et al. 2016 "Practice makes perfect..." (Tsodyks senior author) which identifies the presence of participants using this strategy, in the same data set. Could be good to add a cite to this in some appropriate citation block in the paper.

Signed by Sean Polyn

(Remarks on code availability)

RESPONSE TO REVIEWERS

We thank the reviewers for their insightful feedback, which has been instrumental in guiding the revisions and improvements to our manuscript.

This document includes detailed responses to each reviewer comment. Editor and reviewer comments are shown in black, our responses in blue, and quotes from the revised manuscript in green.

Reviewer 1

Summary

> The authors present a simulation analysis of a set of bespoke computational models trained to optimize performance in free recall tasks. They find that, in addition to the well-documented Temporal Context strategy, characterized by relatively symmetric conditional recall probabilities, some of the networks, under certain conditions, develop what they term a "memory palace"-like strategy.
>
> Overall the findings are very interesting, and generate many testable predictions for real-world experiments on the formation of memory palaces, while also proposing constraints for the utility of this widely celebrated strategy. Our comments focus mainly on tamping down or redirecting the interpretations and claims, and strengthening the connection to both existing TCM theory and empirically observed phenomena.

We sincerely appreciate your positive feedback on our manuscript. Below, we provide a detailed, point-by-point response to each of your comments.

1.1

> The main concern I have is that the way temporal context is operationalized here leaves out something important in how it has been conceived in previous studies. Drifting temporal context doesn't just emerge entirely from internal states – it also consists of representations of (temporally evolving) external stimuli (e.g. physical location, temporal/rhythmic fluctuations in light and background noise). What would happen if you include this sort of information in the previously one-hot encoded input (usually this is done as a vector concatenated to the stimulus representation)?

Thank you for this important point. We agree that temporal context in natural settings can include externally varying information, not only internally generated state dynamics. Earlier models of free recall used random vectors as context (Mensink and Raaijmakers, 1989; Murdock, 1997), and later the temporal context model used retrieved context of studied items to guide memory recall (Howard and Kahana, 2002). In more naturalistic settings, both random environmental information and the drifting context associated with studied items can contribute to temporal context. We therefore agree that it is important to include external information as part of the model input.

To test this directly, we added a slowly drifting vector in the shape of a Gaussian curve as the external context input, which was concatenated with the original item input (Figure 6A). We trained 20 models with different random seeds on context amplitudes of 0, 0.2, 0.4 and 0.8. We used a training setting of discount factor = 1.0, and working memory flushed between phases = 0.9 to make the group of models include various strategies before modulating the amplitude of external context, which is the same as the setting we used for studying the performance of models in section 2.4. Details of methods is specified at line 556.

We found that stronger external context shifted models away from the memory palace strategy and toward more TCM-like strategies (Figure 5, copied as Figure R1 here). Specifically, as context amplitude increased,

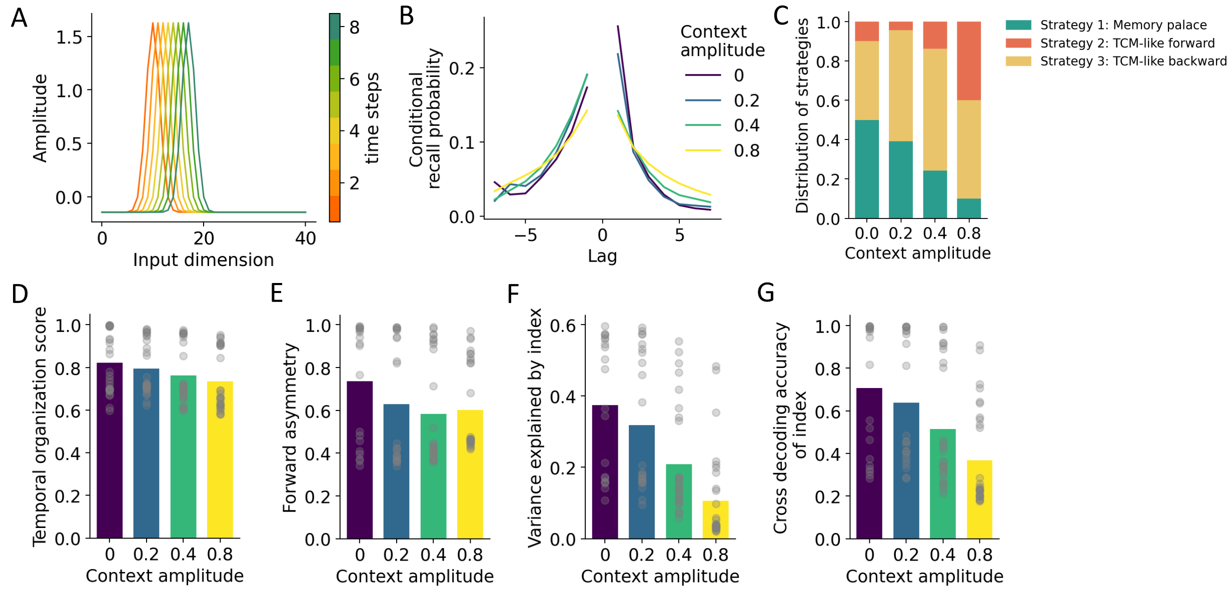

Figure R1: **Change of strategy when external context is included.** (A) An example sequence of context vectors as inputs during the study phase. The vector is a 40 dimensional gaussian bump. Different colors represent input vectors at each time step. For each time step, the vector rolls a dimension over, which results in a slowly drifting temporal context. For each trial, the peak of the context vector at the first time step is randomly selected. (B) Mean conditional recall probability curves of models trained with different amplitudes of context. 20 models with different random seeds were trained for each setting. To make the group of models show various strategies, we used the same training setting as Figure 4 (discount factor = 1.0, proportion of working memory flushed = 0.9). (C) Proportion of models showing each strategy when trained with different amplitudes of context. Strategies were clustered using the same features and clustering centers as in previous figures. (D) Mean temporal organization score, (E) forward asymmetry, (F) Explained variance of index in the hidden states, (G) Cross-decoding accuracy of index from the hidden states for models trained with different amplitudes of context. Grey dots represent each single model.

the mean conditional recall probability curve became less sharp (Figure R1B), fewer models adopted the memory palace strategy and more adopted TCM-like strategies (Figure R1C), temporal organization score and forward asymmetry both decreased (Figure R1D,E), and index-based representation was reduced, both in explained variance and cross-decoding accuracy (Figure R1F,G). We have added these results to the manuscript at line 253:

> The preceding analyses used one-hot item inputs with no environmental information. However, in more realistic settings, temporally varying external signals—such as background sensory input—can serve as an additional source of temporal context. We therefore asked whether introducing such external information would alter the strategies learned by the model. To test this, we concatenated a slowly drifting Gaussian bump vector to the item input as external context (Figure 5A). This signal was independent of item identity and drifted by one dimension per time step. We trained models across a range of context amplitudes and found that stronger external context progressively shifted models away from the memory palace strategy: conditional recall probability curves became flatter (Figure 5B), more models adopted TCM-like strategies (Figure 5C), and temporal organization scores, forward asymmetry, and index representations all decreased (Figure 5D–G). These results indicate that when external temporal context is available as an input, models rely on it rather than learning an internal positional scaffold.

1.2

> The fact that flushing WM leads to memory palace – suggests that to maximize the efficacy of this strategy people should study in different environments than they retrieve, a counterintuitive but exciting finding. Is this finding robust to the inclusion of a more traditional context vector?

Thank you for this interesting suggestion. We want to clarify that flushing working memory in our model is better interpreted as simulating a delay or distractor between study and recall, not necessarily a change in environment. The environment could remain the same even when working memory is disrupted.

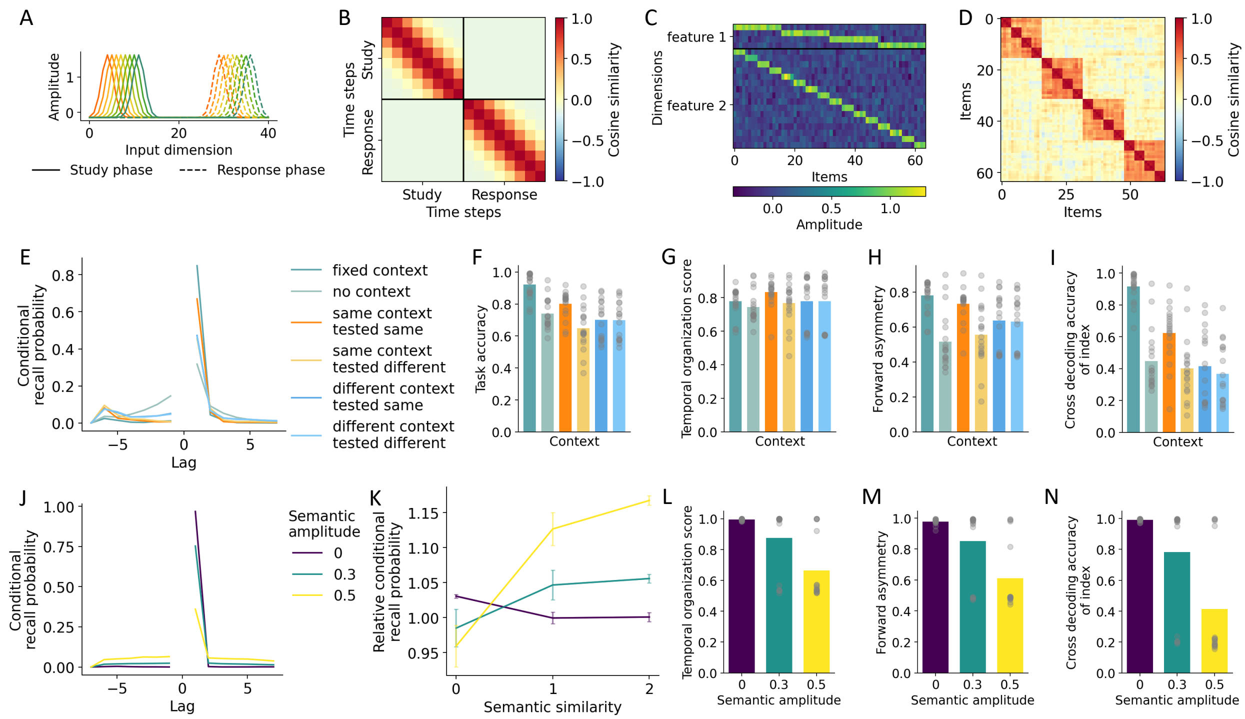

Figure R2: **Change of strategy with temporal and semantic context included as inputs.** (A) Example context vectors as inputs when different context is provided in the study phase and in the response phase. Context within a phase drift slowly, while context across phases are separated. (B) Cosine similarity between context provided at different time steps, when context is different in the study phase and in the response phase. Context within a phase has higher similarity when provided closer in time, while context across phases are dissimilar. (C) Semantic vector for each item as inputs. There are 2 features for each item. The first feature is 4 dimensional, and every 16 items share the same feature. The second feature is 16 dimensional, and every 4 items share the same feature. Gaussian noise is added to the vectors. (D) Cosine similarity between the semantic vectors for items. There are 3 levels of semantic similarity. (E) Mean conditional recall probability for models with the same context provided in the study phase and the response phase, and models with different context provided across phases. 20 models with different random seeds were trained for each setting. Each model was tested with the same context or different context provided across phases. (F) Mean free recall performance of models trained with the same or different context across phases, and tested with the same or different context, respectively. (G) Temporal organization score, (H) forward asymmetry, (I) cross-decoding accuracy of index of models trained and tested with the same or different context. (J) Mean conditional recall probability for models trained with different amplitude of semantic information as inputs. 20 models with different random seeds were trained for each setting. (K) Mean relative conditional recall probability as a function of semantic similarity level for models trained with different amplitude of semantic information. (L) Temporal organization score, (M) forward asymmetry, (N) cross-decoding accuracy of index of models trained with different amplitude of semantic information.

That said, the relationship between encoding/retrieval context is an important question, so we directly tested it. We trained models with either matching or mismatched external context across study and response phases (Extended data figure 6A,B, here copied as Figure R2). We found that models trained with matching context performed better and were more likely to learn the memory palace strategy than those trained with mismatched context (Figure R2E-I). Additionally, when models trained with matching context were tested with mismatched context, both performance and memory palace adoption decreased. The reverse—training with mismatched context but testing with matching context—produced no significant benefit.

These findings align with the well-established context-dependent memory literature (Smith and Vela, 2001) and suggest that consistent context across encoding and retrieval may function as location-like information, guiding the model toward a memory palace strategy. To test this hypothesis, we trained the model with the same sequence of context vectors across both trials and phases. In this setting, we found that the model had a very high tendency to learn index code, exceeding that of models trained with the same context across phases and models trained with no context.

The results are reported at line 264, with methods at line 561.

1.3

> Were other naturalistic strategies possible given these impoverished stimulus classes? e.g. clustering by conceptual/featural similarity? (I remember heart suit cards mostly together, etc) One way to examine this may be to ask whether similar stimuli (e.g. by cosine distance) are retrieved together. If not, is there a reason that these strategies never developed? Might they have developed with more rich stimuli that had been processed more categorically by something like the visual processing hierarchy in brains?

This is an excellent point. Our original stimuli were one-hot vectors with no shared features, which precluded similarity-based clustering. To test whether richer stimuli would elicit such strategies, we trained models with a two-level hierarchical feature embedding (Figure R2C,D): each item had two features, one shared across every 4 items and another shared across every 16 items, creating three levels of semantic similarity. Methods are at line 566.

We found that, as semantic information increased in amplitude, models showed reduced temporal contiguity (Figure R2J) and an increased tendency to recall semantically similar items in succession (Extended data figure 6K). Models also became less likely to adopt the memory palace strategy (Figure R2L-N). This is consistent with findings that strong semantic structure reduces temporal contiguity effects (Healey and Uitvlugt, 2019; Hong et al., 2024). These results are reported at line 277.

1.4

> Fig 1G seems to operationalize the distinction between temporal context and memory palace in that the latter should show no backward recall. Is this strict forward processing actually observed in real memory palace-driven behavior? Do individuals not sometimes "retrace their steps?" From a quick search there appears not to be much laboratory work on this, but videos of memory champions suggest that it happens at least sometimes – indeed it appears to be encouraged by some practitioners as a way of building a robust "palace."

We agree that strict forward-only recall is an idealization, not a requirement, of the memory palace technique. The key feature is associating items with an ordered sequence of mental locations, which creates a strong forward tendency but does not preclude backtracking (Yates, 2013; Worthen and Hunt, 2011). Indeed, Roediger, 1980 found that memory palace particularly benefits tasks requiring recall in encoding order, but backward traversal and error correction are clearly part of expert practice. Furthermore, simply observing a high forward recall tendency cannot predict whether the behavior is based on temporal contexts or positional codes, as optimal TCM (Zhang et al., 2023) also leads to a forward recall behavior.

In our simulations, most memory palace models showed very high—but not perfect—forward asymmetry, and we observed a strong positive correlation between forward recall tendency and index code learning (Figure 3D). We therefore characterize forward recall as a hallmark property of the memory palace strategy rather than a strict requirement.

We have revised the manuscript (section 2.1) to reflect these important points.

1.5

> Relatedly, do humans ever exhibit such extreme forward-only recall in free recall tasks, absent the explicit instruction to start from the beginning, given ahead of study?

Yes. We analyzed the PEERS dataset (Kahana et al., 2024), which includes 200 young adults performing immediate free recall (list length 16) with no special recall instructions. We identified participants matching each of the three model strategies. A small subset showed very high forward recall tendencies (Extended data figure 3A), though with some backward recall. These participants were not trained to use any specific technique. We also found that forward recall tendency in humans was positively correlated with free recall performance (Extended data figure 5), paralleling our model results.

1.6

> It is commonly thought that memory palaces are supported by spatial maps. Did the authors observe evidence in the internal activations of the trained networks that are consistent with map-like representations? Alternatively, are the representations instead better thought of as directed line-graphs? Would more naturalistic representations yield memory palaces under different conditions?

Great question. Since our task contains no spatial location information, we would not expect the model to develop two-dimensional map-like representations. The index representation that emerged is best described as a one-dimensional directed line-graph closer to what you suggest, where each serial position serves as a distinct and stable location for item binding. We treat this as an abstract form of the "loci" used in memory palace, consistent with evidence that an explicit spatial environment may not be essential as long as stereotyped location information is available for item binding (Caplan et al., 2019), and temporal information such as a set of detailed autobiographical events can also guide learning of memory palace (Bouffard et al., 2018).

Regarding more naturalistic representations: in our new simulations incorporating external context, we found that consistent context during encoding and retrieval can serve as location information that guides the model toward memory palace learning. If this were replaced with richer spatial representations, we would expect the model to develop a more human-like memory palace, binding items to locations in a structured mental environment with rich details.

1.7

> Similarly, the idea of memory palaces usually involves scaffolding on existing, well learned, often personally meaningful spatial arrangements, not just a simplistic ordinal position. Can the authors discuss some of the potential differences between their ordinal code and the one they observed here? Propose an architecture that would be able to leverage this kind of knowledge?

We appreciate this suggestion. The index code in our model is a highly abstract form of loci, capturing only serial position. Human memory palace relies on a far richer scaffold, such as a two-dimensional spatial environment grounded in personal experience, enriched with vivid imagery to strengthen item-location associations (Huang et al., 2024).

Several existing architectures point toward bridging this gap. Whittington et al. (2020) proposes a model that learns relational memory by analogy to spatial map formation, and Chandra et al. (2025) demonstrates memory palace capability by abstracting memory into low-dimensional spatial transitions. However, neither model focuses on the temporal retrieval dynamics central to free recall. A natural extension would be to integrate such spatially-grounded architectures with our framework, enabling the study of how temporal retrieval properties interact with richer spatial scaffolds. We discuss this at line 343:

> The index code learned by our model can be understood as an abstract form of loci: each serial position in the list serves as a distinct, stable location onto which items are bound during encoding and systematically retraced during retrieval. In practice, humans using a memory palace technique often rely on a richer two-dimensional spatial scaffold grounded in personally meaningful environments, where perceptual detail and vivid imagery further strengthen item-location associations. A natural extension of our work would be to integrate our framework with architectures that learn relational memory by analogy to spatial map formation (Whittington et al., 2020; Chandra et al., 2025). This could yield a spatially-grounded architecture that preserves the temporal retrieval dynamics while also supporting the richer associative structure that characterizes expert-level human memory performance.

Reviewer 2

Summary

Li and colleagues present a creative study of human memory search that bridges with the machine intelligence literature in a novel and valuable way. In my mind, this paper makes a substantial contribution to both cognitive psychology theory and machine intelligence theory. In theories of human memory, the free-recall task provides valuable empirical data regarding the dynamic processes underlying memory search. One of the foundational computational models in this domain is the TCM model described in the introduction and used as a reference point throughout the paper. TCM is one of a class of retrieved-context models, and described many important behavioral phenomena, including the phenomenon at the center of this paper, regarding the temporal organization of human memory. In a sense, TCM is an example of a hand-crafted neural network model, designed to be simple, with very few free parameters. This makes the mechanisms of the model more interpretable, but also can limit the flexibility of the model. A major contribution of the current paper is showing that modern RNN technology can effectively learn to perform memory search in a variety of ways, reflecting the different ways human participants are known to approach memory search.

Making bridges between the modern technologies of machine intelligence, and the decades-long history of theoretical development of cognitive memory theory will have benefits for both domains. Here, the authors demonstrate that their novel modeling framework can learn to produce multiple kinds of internal representational structures and behaviors seen in humans. These span from serial recall with their connection to the memory palace technique, and to free recall, with 2 forms of behavior evident (forward biased and backward biased transitions).

The analysis of positional vs temporal internal states is excellent and has the potential to bridge two historically distinct lines of theory development in cognitive and mathematical psychology.

The addition of a memory module to the GRU system is innovative, in that a set of hidden states are formed during the simulated study period, much like the stored representations in a modern Hopfield model, or more aptly for the cognitive psychological memory literature, Hintzman's MINERVA model which precedes modern Hopfield by decades but has the same basic architecture and purpose.

> The final section examining how an alternative recall task can alter the optimal strategy is also innovative and theoretically important. It is understandably shorter than the other sections and I look forward to seeing it expanded in future work.

Thank you for your positive and thoughtful assessment. We have incorporated your suggestions in the revised manuscript as detailed below.

2.1

> Given the emphasis on human behavior, the paper would benefit from a figure showing examples of the different free recall strategies pulled from existing empirical data sets. Unless I'm missing something, all the strategy examples are pulled from the network behavior. Could have a separate figure demonstrating empirical data showing these things. I imagine the PEERS data set has all 3 in spades across participants. I don't think it would be necessary to integrate the empirical data into the existing analysis framework though, it seems cleaner as is, showing that these patterns arise naturally from the task/training demands.

Thank you for your suggestion. All the strategy examples in Figure 2 were indeed from model simulations. To connect our findings more directly to human behavior, we analyzed the PEERS dataset (Kahana et al., 2024), extracting behavioral data from young adults performing immediate free recall with list length 16. We grouped participants by forward asymmetry and temporal organization score (methods in section 4.7, line 544).

We identified participants whose behavior matched each of the three model strategies, now shown in Extended data figure 3 (copied here as Figure R3). This is described at line 191:

> To compare the three RNN strategies with human free recall, we analyzed data from the Penn Electrophysiology of Encoding and Retrieval Study (PEERS), where young adults performed an immediate free recall task (Kahana et al., 2024). By analyzing forward asymmetry and temporal organization scores, we identified a subset of participants who exhibited a strong tendency to recall items in forward order, consistent with the memory palace strategy (Extended data figure 3A). Others displayed smoother temporal contiguity effects with forward asymmetry (Extended data figure 3B), while a third group showed backward asymmetry (Extended data figure 3C). These data suggest that the three strategies identified in our model also reflect naturally occurring recall patterns in humans.

We also found that both forward asymmetry and temporal organization score were positively correlated with free recall performance in humans (Extended data figure 5, copied here as Figure R4), consistent with previous findings (Sederberg et al., 2010) and paralleling our model results. This is reported in section 2.4 at line 230:

> Notably, we found a similar trend in the human data. In the PEERS dataset, both forward asymmetry and temporal organization score were positively correlated with free recall performance(Extended data figure 5A,B).

2.2

> Section 4.6 in the methods, I can't find mention to this in the main paper, though it seems a quite valuable description making an explicit connection between this work and modern transformer theory. I recommend at least adding some forward references to this section in the main body of the paper, if not moving some version of the section to earlier in the paper.

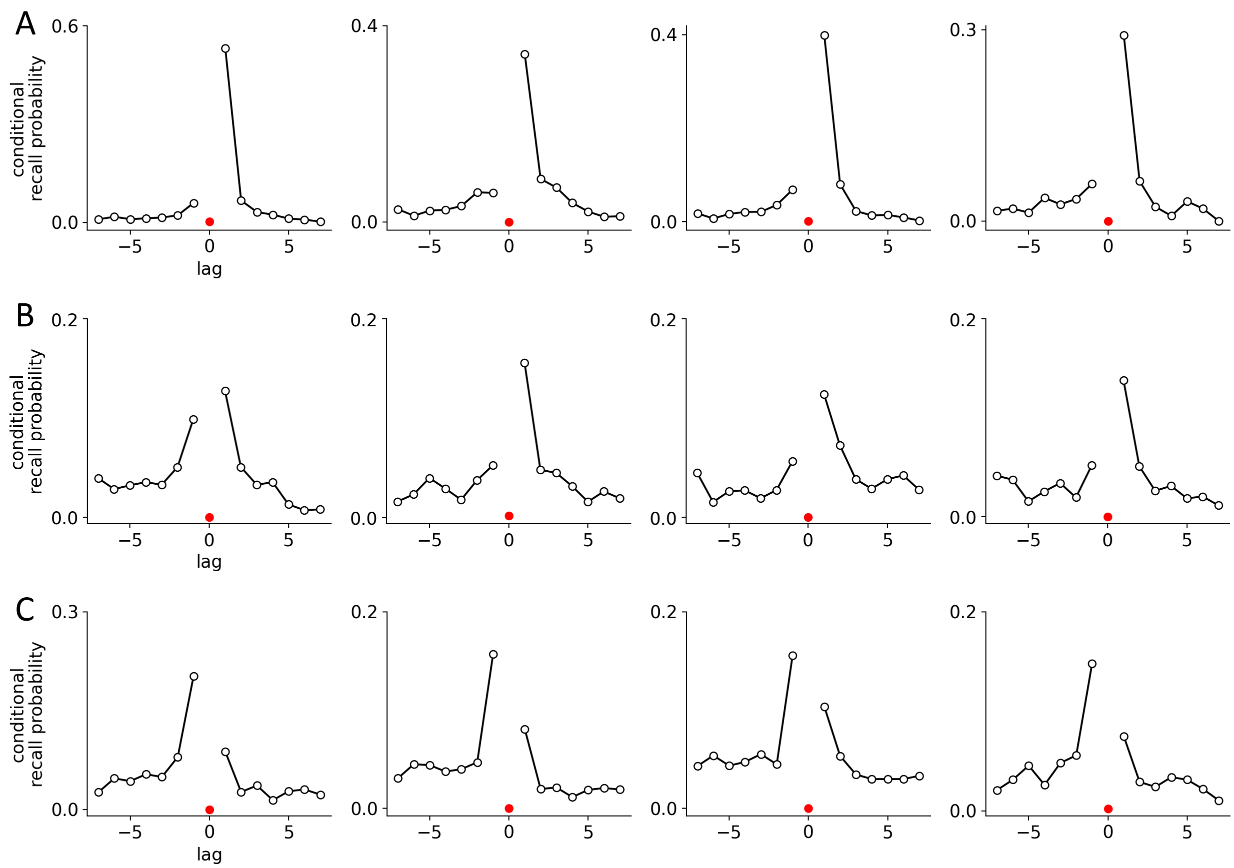

Figure R3: **Behavior of human participants match to each strategy.** Four human participants from the PEERS dataset were selected to match the behavior of each strategy. Participants were manually selected through filtering with forward asymmetry and temporal organization score. (A) Conditional recall probability of participants that match the memory palace strategy. (B) Conditional recall probability of participants that match the TCM-like forward strategy. (C) Conditional recall probability of participants that match the TCM-like backward strategy.

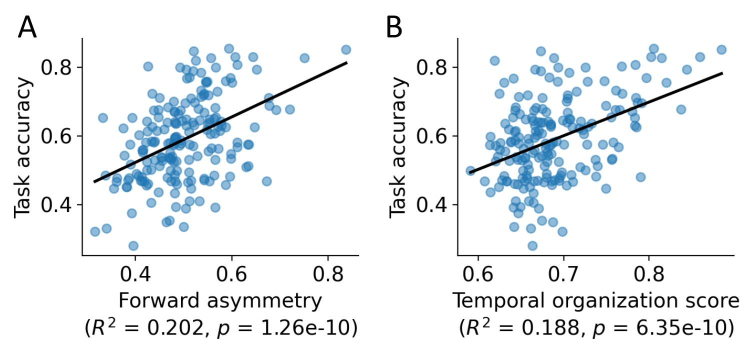

Figure R4: **Relationship between behavioral metrics and free recall performance in human data.** Relationship between (A) Forward asymmetry and (B) temporal organization score and the task performance for 200 participants in the PEERS dataset.

Thank you for catching this omission. This simulation shows that the memory palace strategy also emerges with a more general key-value memory architecture widely used in recent machine learning work on episodic memory. We have added a description in section 2.2, at line 137:

> The memory palace mechanism also emerged when using a more general key-value memory module that supports distinct representations for encoding memory location (keys) and content (values) (Extended data figure 1A,B; Zhang et al., 2017; Ritter et al., 2018). This generalized architecture can account for some features observed in the human memory system (Gershman et al., 2025; Song et al., 2025). When analyzing the variance in keys and values explained by item index and item identity, we found that keys more strongly represented item index, while values more strongly represented item identity (Extended data figure 1C). This functional separation may play an important role in the formation of the memory palace strategy in the human brain, which could similarly segregate index-based and identity-based representations during structured recall.

2.3

> Would be interesting to hear a brief thought in conclusions about extending the model to situations where the model components are not wiped clean between trials.

This is an important point. In our current model, memory is cleared between trials. In a multi-trial setting, the memory palace strategy could face interference when different items are bound to the same positional locations across trials. TCM-like strategies, which incorporate event context, may be better suited for distinguishing items across episodes. We have added this to the discussion at line 336:

> Furthermore, our model clears its memory across trials, which does not account for situations that require multiple separate sequences to be kept in memory. Humans distinguish between different events using event context. This could make the memory palace less optimal, since items stored at the same location may interfere with one another in the absence of an event context.

2.4

> Figure 2A caption. "Colors indicate strategy groups" – clarify what each colored dot represents. "dots show reservoir RNN samples" – does this mean black dots? There are colored and black dots, could clarify for the reader.

We have revised the caption for clarity:

> (A) tSNE embedding of clustering metrics from all trained networks. Different colors denote clusters of models exhibiting strategies 1, 2, and 3 below. (B) Relationship between forward asymmetry and temporal organization score. Black dots indicate reservoir networks, and the other colors indicate trained models exhibiting each of the three strategies.

2.5

> page 13. Temporal factor. Small correction: This technique was introduced by Polyn, Norman & Kahana Psych Review (2009) described on page 38. It was used the following year by Sederberg et al. (2010).

Thank you for this correction. We have updated the citation at line 475:

> The temporal organization score (TS, also named as temporal factor) was introduced by Polyn et al. (2009), and is a widely used metric that quantifies the tendency of an agent to recall an item that is in close temporal proximity to the previously recalled item (Sederberg et al., 2010; Hong et al., 2024).

2.6

> Temporal factor name. This analysis was initially called temporal factor, but in more recent years some groups have shifted to calling it a 'temporal organization score' or the more unwieldy 'percentile rank temporal organization score'. Temporal factor is more succinct, but there's a minor negative in that it can be misinterpreted as a form of factor analysis. Hong et al. (2023) JEP:LMC (The modulation and elimination of temporal org in free recall) provides a detailed description of the algorithm and relates it to the lag-CRP algorithm also used in this paper.

We agree that "temporal organization score" is less prone to misinterpretation and have adopted it throughout the manuscript in place of "temporal factor." We have also cited Hong et al. (2024) in the methods section where the metric is described.

References

Bouffard, N., Stokes, J., Kramer, H. J., and Ekstrom, A. D. (2018). Temporal encoding strategies result in boosts to final free recall performance comparable to spatial ones. *Memory & Cognition*, 46(1):17–31.

Caplan, J. B., Legge, E. L., Cheng, B., and Madan, C. R. (2019). Effectiveness of the method of loci is only minimally related to factors that should influence imagined navigation. *Quarterly Journal of Experimental Psychology*, 72(10):2541–2553.

Chandra, S., Sharma, S., Chaudhuri, R., and Fiete, I. (2025). Episodic and associative memory from spatial scaffolds in the hippocampus. *Nature*, 638(8051):739–751.

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

Healey, M. K. and Uitvlugt, M. G. (2019). The role of control processes in temporal and semantic contiguity. *Memory & Cognition*, 47(4):719–737.

Hong, M. K., Gunn, J. B., Fazio, L. K., and Polyn, S. M. (2024). The modulation and elimination of temporal organization in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 50(7):1035.

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

Kahana, M. J., Lohnas, L. J., Healey, M. K., Aka, A., Broitman, A. W., Crutchley, P., Crutchley, E., Alm, K. H., Katerman, B. S., Miller, N. E., et al. (2024). The penn electrophysiology of encoding and retrieval study. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Mensink, G.-J. M. and Raaijmakers, J. G. (1989). A model for contextual fluctuation. *Journal of Mathematical Psychology*, 33(2):172–186.

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (todam2). *Psychological Review*, 104(4):839.

Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological review*, 116(1):129.

Ritter, S., Wang, J., Kurth-Nelson, Z., Jayakumar, S., Blundell, C., Pascanu, R., and Botvinick, M. (2018). Been there, done that: Meta-learning with episodic recall. In *International conference on machine learning*, pages 4354–4363. PMLR.

Roediger, H. L. (1980). The effectiveness of four mnemonics in ordering recall. *Journal of Experimental Psychology: Human Learning and Memory*, 6(5):558.

Sederberg, P. B., Miller, J. F., Howard, M. W., and Kahana, M. J. (2010). The temporal contiguity effect predicts episodic memory performance. *Memory & cognition*, 38:689–699.

Smith, S. M. and Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic bulletin & review*, 8(2):203–220.

Song, H., Lu, Q., Nguyen, T. T., Chen, J., Leong, Y. C., Rosenberg, M. D., Ching, S., and Zacks, J. M. (2025). A neural network with episodic memory learns causal relationships between narrative events. *bioRxiv*, pages 2025–09.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T. E. (2020). The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263.

Worthen, J. B. and Hunt, R. R. (2011). *Mnemonology: Mnemonics for the 21st century*. Psychology Press.

Yates, F. A. (2013). *Art of memory*. Routledge.

Zhang, J., Shi, X., King, I., and Yeung, D.-Y. (2017). Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th international conference on World Wide Web*, pages 765–774.

Zhang, Q., Griffiths, T. L., and Norman, K. A. (2023). Optimal policies for free recall. *Psychological Review*, 130(4):1104.

RESPONSE TO REVIEWERS

Reviewer 1

1.1

> I thank the reviewers for their comprehensive response to my concerns.

We sincerely appreciate your positive feedback on our manuscript.

Reviewer 1

2.2

> The authors did an outstanding job responding to the reviews. I believe the manuscript is much strengthened as a result, and will make an excellent contribution to the literature.

We sincerely appreciate your positive feedback on our manuscript.

2.3

> The only minor thing I noted; with regard to the inclusion of the PEERS data in the extended figure identifying participants that indeed show an extreme version of the forward chaining / memory palace behavior in the lag-CRP. There is a paper by Romani et al. 2016 "Practice makes perfect..." (Tsodyks senior author) which identifies the presence of participants using this strategy, in the same data set. Could be good to add a cite to this in some appropriate citation block in the paper.

Thank you for your suggestion. We have added this reference to our paper where we reported human participants showing the extreme forward recall behavior:

By analyzing forward asymmetry and temporal organization scores, we identified a subset of participants who exhibited a strong tendency to recall items in forward order, consistent with the memory palace strategy (Extended data figure 3a) (Romani et al., 2016).

References

Romani, S., Katkov, M., and Tsodyks, M. (2016). Practice makes perfect in memory recall. *Learning & memory*, 23(4):169–173.