Tree of Memory: A Mathematical Model of Hierarchical Episodic Recall

Tomer Waizer (tomer.waizer@campus.technion.ac.il)[1],
Stefano Recanatesi (stefano@technion.ac.il)[2] **& Omri Ben-Eliezer (omribene@cs.technion.ac.il)**[1]
[1]Taub Faculty of Computer Science, Technion – Israel Institute of Technology, Haifa, Israel
[2]Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa 3109601, Israel

Abstract

Understanding free recall requires models that explain not only how many items are retrieved, but also how retrieval is *organized*—with clustered output, structured transitions, and strong dependence on temporal position. We introduce the *Tree of Memory (TOM)*, in which recall unfolds as a probabilistic search over an explicitly hierarchical episodic representation, coupled with a minimal semantic association structure. The episodic component represents experience across temporal scales encoded in different hierarchical layers. Recall proceeds via probabilistic depth-first traversal of the episodic tree, interleaved with semantic exploration triggered upon item retrieval. The framework admits explicit mathematical characterizations of recall-capacity scaling regimes: depending on the scaling of attention-weighted episodic accessibility, the expected number of recalled items grows sublinearly with list length. In simulations, TOM reproduces canonical free-recall signatures, thereby providing a principled framework to study how hierarchical episodic organization, attention, and semantic associations jointly shape free recall.

Keywords: episodic memory; free recall; list-length effect; attention; hierarchical models; semantic memory; mathematical modeling

Introduction

Modeling human episodic memory is a central challenge in cognitive science, precisely because memory exhibits both vast capacity and strikingly systematic patterns of retrieval across a wide range of tasks. Classical frameworks such as the Search of Associative Memory (SAM) model (J. Raaijmakers & Shiffrin, 1981), the Temporal Context Model (TCM) (Howard & Kahana, 2002) and others (Kahana, 2020; J. G. W. Raaijmakers & Shiffrin, 2002; Shiffrin & Steyvers, 1997) have been highly influential, successfully accounting for many qualitative and quantitative regularities in free recall (Kahana, 1996; B. Murdock, 1960; Shepard, 1967). These models, and their numerous extensions, typically embody rich stochastic dynamics and a large number of tunable parameters, yielding excellent fits to data but making principled analysis and mechanism isolation difficult. Consequently, many of their key predictions are evaluated through extensive simulations rather than through explicit analytic characterizations.

Motivated by the desire to uncover first principles of recall organization, a parallel line of work has explored simplified, mathematically characterized models of free recall. Recent approaches in this direction reduce recall to a search (or walk) over a structure induced by memory representations, yielding compact predictions for the average number of recalled items as a function of list length and matching large-scale datasets (Katkov et al., 2022; Naim et al., 2020). These results clarify that recall is strongly capacity-limited in a lawful, sublinear way. However, many analytically tractable accounts intentionally abstract away explicit multi-level episodic organization, even though hierarchical structure is a pervasive feature of memory and can markedly improve recall when present (Bower et al., 1969). Tree-based hierarchical models have also recently been used to formalize recall and compression in *meaningful* narrative memory (Zhong et al., 2025) and retrieval (Zhong et al., 2026). Moreover, hierarchical organization can emerge even for nominally unstructured material, suggesting that multiscale episodic structure is not merely a feature of meaning, but a general organizing principle of retrieval (Naim et al., 2019). This motivates our central question:

> *Can an explicit episodic hierarchy account for canonical free-recall effects while remaining analytically predictive?*

In this work, we introduce the *Tree of Memory* (TOM), a framework that models free recall as a probabilistic search over a *hierarchically structured episodic representation* coupled with semantic associations. The episodic component is a tree in which each level represents experiences at a different temporal scale. Lower levels correspond to short-timescale events (e.g., individual words), while higher levels represent progressively longer-timescale episodes (e.g., groups of successive words). Although TOM is not restricted to a fixed-height tree, in the single-list free recall setting we focus on the local subtree rooted at the list episode and consisting of three levels: (i) list episodes, (ii) chunk episodes, and (iii) item episodes. This instantiation reflects empirical evidence that people group items into small clusters during encoding and retrieval (Cowan, 2001; Farrell, 2012).

A key modeling choice in TOM is that episodic accessibility varies systematically across positions within each episodic level. Concretely, each episodic edge is assigned an attention-weighted traversal probability: letting e_n denote the n-th edge created at a given level, its traversal probability is $P(e_n) = f(n)$ where $f : \mathbb{N} \to [0, 1]$ is monotone non-increasing. Earlier edges are more likely to be followed, creating a natural primacy bias. During recall, episodic retrieval is implemented as a *probabilistic depth-first search* (DFS) over the episodic tree. In the single-list paradigm, recall is initiated at the list node, consistent with the assumption that the most recent episodic context is directly accessible at test.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

TOM also includes a semantic component that represents relationships between items based on meaning. For simplicity, we consider two minimal semantic structures (i.e., an Erdős–Rényi graph and a geometric embedding) that are not intended as complete models of semantic memory, but rather as controlled ways to study how semantic associations interact with episodic retrieval. Operationally, semantic retrieval is triggered upon recalling an item and proceeds as a local DFS-like exploration over semantic edges; once semantic exploration is exhausted, control returns to the episodic DFS.

We evaluate TOM in the single-list free recall paradigm and show that it reproduces several hallmark phenomena, including the primacy effect in the serial position curve (B. B. Murdock, 1962), temporal contiguity (lag-CRP) (Kahana, 1996), and the list-length effect. In addition, TOM admits explicit recall-capacity scaling regimes with list length k: depending on the magnitude and scaling of the attention-weighted traversal probabilities (as controlled by f), the expected number of recalled items can grow either logarithmically or as a sublinear power law. Simulations validate these predictions and illustrate how episodic and semantic retrieval act as complementary mechanisms. Although the current model does not directly capture recency, it can be naturally extended, in the spirit of SAM (J. Raaijmakers & Shiffrin, 1981), by incorporating a constant-capacity short-term store for recently studied items.

More broadly, TOM is intended as a step toward models that place *hierarchical episodic organization* at the center of representation and retrieval, while remaining sufficiently constrained to yield quantitative predictions for standard free recall.

TOM Definition

The *Tree of Memory (TOM)* is designed to capture how people organize and retrieve memories during free recall. The model represents memory as a hierarchy of *episodes* occurring at different temporal scales. Each node in the tree corresponds to a memory episode at a particular timescale, ranging from coarse summaries of experience to fine-grained representations of individual events. The model has two interacting components. The first is a tree-structured episodic representation that encodes when items were experienced across multiple timescales. The second is a semantic structure that represents relationships between items based on meaning. Together, these components capture temporal and semantic organization in recall. Because our primary focus is on the episodic component, we adopt two simple semantic structures that are not intended to provide a fully realistic account of semantic memory, but rather to illustrate how semantic associations can interact with episodic retrieval.

In the episodic component, higher levels of the tree correspond to broader temporal episodes that summarize longer stretches of experience, while lower levels represent shorter, more fine-grained episodes, as shown in fig. 1. In our setting, the levels of the tree correspond to:

1. **List episodes**: individual study lists within the session,
2. **Chunk episodes**: short sequences of successive items within each list,
3. **Item episodes**: the individual words at the leaves of the tree.

In general, the model is not restricted to a tree of fixed height; longer temporal spans can be represented by introducing additional hierarchical levels. For the list-learning experiments considered here, however, we focus on a local subtree consisting of the bottom three levels rooted at the list node. Although this list node may itself be embedded within a larger hierarchy corresponding to earlier experiences, we omit these higher levels from the analysis. This is justified because the list is the most recent episode, so retrieval begins from the list node. Empirically, people often group items into small clusters of 2–6 elements during free recall (Cowan, 2001; Farrell, 2012), and our model reflects this by giving each chunk a fixed small capacity C.

For the semantic component, we aim for simplicity over realism, and consider two possible representations. One is an Erdős–Rényi graph (Erdős & Rényi, 1960), in which each node is a word and edges link semantically related items. The other is a high-dimensional sphere on which words are embedded such that nearby points correspond to semantically similar items. This semantic structure is treated as fixed throughout the experiment, while the episodic tree grows dynamically as words are presented.

Constructing TOM

The episodic tree is constructed online during study. At the onset of a list, a new list node is attached to the most recent experience, represented by a higher-level node in the tree corresponding to a longer timescale episode of which the list presentation is a part. In the single-list free recall paradigm, we focus only on the subtree rooted at the list node, since recall is initiated immediately after list presentation and the list context is therefore readily accessible. In contrast, if additional tasks intervene before recall, the list node may no longer be directly accessible, and recall must instead be initiated from a higher episodic level. Each incoming word w_t is assigned to the most recently created chunk, which is a child of the list node, unless that chunk has reached its capacity. In that case, a new chunk is created as an additional child of the list node, and w_t becomes its first child.

Attention-weighted edge probabilities. To capture the observation that attention tends to wane over time, we associate each edge in the tree with a probability of successful traversal during recall. Let e_n denote the n-th edge created at a particular level of the tree. We call the probability of successfully traversing that edge

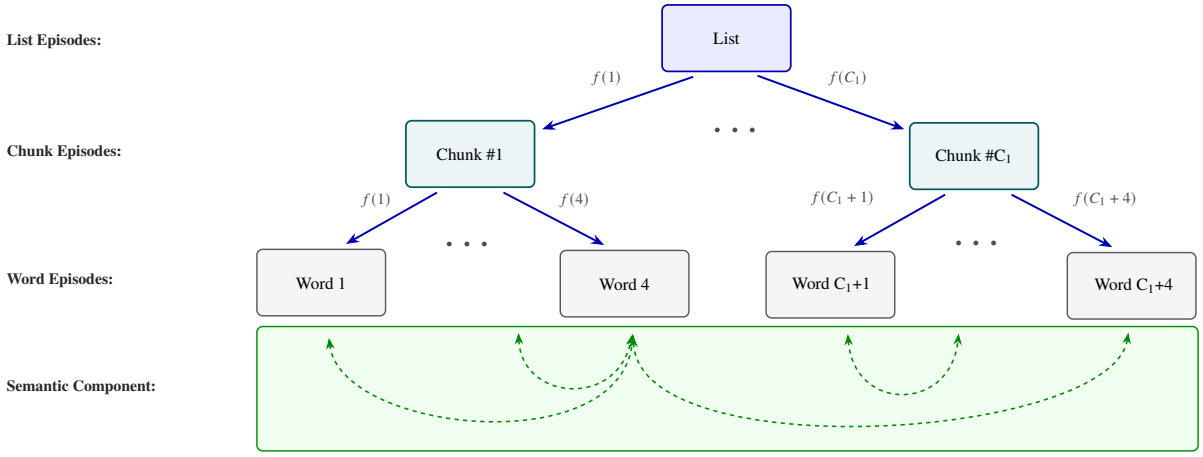

Figure 1: TOM with Random Semantic Connections

the *attention-weighted edge probability*, and denote it, as in fig. 1 by

$$P(e_n) = f(n),$$

where $f : \mathbb{N} \to [0, 1]$ is a monotonically non-increasing function.[1] Earlier edges therefore receive higher traversal probabilities, naturally giving rise to a *primacy advantage*: items encoded earlier in a list are more accessible at retrieval.

Recall Procedure

Recall is modeled as a probabilistic depth-first search (DFS) through the episodic tree, augmented by occasional semantic transitions between words. Recall unfolds as a sequential exploration alternating between episodic structure and semantic associations.

Recall dynamics. Recall in the model proceeds via two interacting mechanisms: episodic recall and semantic recall. We describe each mechanism separately and specify how control passes between them.

Episodic recall. Episodic recall is modeled as a probabilistic depth-first search over the episodic tree.

1. Recall is initiated at a designated episodic node. In a single-list free recall paradigm (fig. 1) this corresponds to the *List* node.
2. If the current *episodic node* v is a leaf of the episodic tree (i.e., a *Word*), the item stored at v is recalled, and control is passed to the *semantic recall* system.
3. Otherwise, the model considers all outgoing episodic edges from the current node v to its children. The model then picks a random ordering over the outgoing edges from v by sampling these edges without replacement; when creating this ordering, the probability that each edge is selected at any time is proportional to its attention-weighted traversal probability.[2]
4. For each selected episodic edge e, the model attempts to traverse it (according to the attention-weighted probability associated with e). If traversal succeeds, the model moves to the corresponding child node and repeats this procedure recursively. If traversal fails, the edge is discarded and the model attempts another outgoing edge from the current node.
5. When no further outgoing episodic edges remain at the current node, episodic traversal from that node terminates and the model backtracks to its parent.

In step 3 of episodic recall, given a node v, the model samples outgoing edges with probability proportional to their attention-weighted traversal probabilities. That is, for $u \in N(v)$,[3] we have

$$\Pr[\text{sample } u] = \frac{\Pr[\text{traverse } u]}{\sum_{w \in N(v)} \Pr[\text{traverse } w]}.$$

Semantic recall. Semantic recall is initiated whenever a word is recalled during episodic traversal. It is also based on a depth-first search, but unlike the episodic setting, the underlying structure is a graph.

1. Upon entering semantic recall, the model considers the outgoing semantic edges from the recalled word.
2. Semantic edges are traversed sequentially, and each successful traversal leads to the recall of a semantically associated word.
3. From each newly recalled semantic word, semantic traversal continues recursively using the same procedure. Semantic recall terminates when no further unrecalled semantic neighbors can be accessed.
4. Once semantic recall is exhausted, control returns to episodic recall at the leaf node corresponding to the initially recalled word, and episodic traversal resumes by backtracking upward.

[1] For simplicity, we use the same attention-weighted function for different levels of the tree. One could of course use different functions for the different levels, without materially changing the model.

[2] This type of ordered sampling is used in (and hence takes inspiration from) popular recall frameworks such as SAM (J. Raaijmakers & Shiffrin, 1981).

[3] $N(v)$ denotes the neighbors of v.

Model Analysis

In this section, we present the main mathematical results of our model, stated as theorems. Our primary and most difficult results address the list-length effect by demonstrating that, for appropriate choices of the model parameters, the expected recall exhibits sublinear scaling with list length.

We first mathematically define a *Free Recall Test*.

Definition 1 (Free Recall Test)**.** A *Free Recall Test* of k words is a procedure in which the model is presented with k consecutive words $w_1, \ldots, w_k$ in serial order. After the presentation phase, the model is asked to recall as many of the words as possible.

Our theoretical results make a mild structural assumption on the semantic component, given below. This assumption holds with high probability[4] assuming the semantic component is modeled using popular random graph models such as Erdős–Rényi $G(k, p)$ with, say, $p \leq 0.9/k$, or a geometric random graph with a similar choice of parameters (Janson & Luczak, 2008; Kupper & Penrose, 2024; Penrose, 2003).

Definition 2 (Subcritical semantic graph)**.** Let k be the number of words in the free recall test. We say that the semantic component over these k words is *subcritical* if, when interpreted as a graph, its connected components[5] $C_1, \ldots, C_m$ satisfy that $\sum_{i=1}^{m} |C_i|^2 = O(k)$.

Our main analytical result shows that our model can achieve realistic recall performance.

Theorem 1 (TOM recall performance)**.** *Consider a free-recall experiment with k studied words and a TOM model with depth d and chunk size C, with a subcritical semantic component. Suppose that all attention-weighted edge probabilities in the model fall in some interval I. Then:*

- *If* $I = \left[\left(\frac{a \log k}{k}\right)^{1/d}, \left(\frac{b \log k}{k}\right)^{1/d}\right]$*, the expected number of recalled words is of order* $\Theta(\log k)$.
- *If* $I = \left[\left(ak^{\alpha-1}\right)^{1/d}, \left(bk^{\alpha-1}\right)^{1/d}\right]$ *for any* $\alpha \in (0, 1)$*, the expected number of recalled words is* $\Theta(k^{\alpha})$.

Theorem 1 shows that TOM admits multiple recall scaling regimes, ranging from logarithmic to sublinear power-law growth. The specific regime realized depends on the magnitude of attention-weighted traversal probabilities. This unifies disparate recall behaviors within a single cognitive architecture and provides a mechanistic account of how variations in attentional decay can give rise to qualitatively different recall capacities. Let us illustrate the application of our theorem with a concrete example.

Consider a free-recall experiment with $k = 256$ studied words. Let $a = \frac{1}{2}$ and $b = 2$. In a single-list recall setting, the model depth is $d = 2$. For $k = 256$, we have $\log(k) = 8$, $\sqrt{k} = 16$. Substituting these values into Theorem 1 (with $\alpha = \frac{1}{2}$) yields the following two admissible intervals for the attention function:

$$I_{\log} = [0.125, 0.25],\ I_{sqrt} = [0.176, 0.353].$$

Theoretical proof sketch

In this section, we present the proof of Theorem 1. The proof proceeds by first analyzing a simplified version of the TOM model in which the attention function is constant, $f(n) = c$ for some $0 < c < 1$. We derive bounds for this *simple-TOM model* and then extend the analysis to the full TOM model.

Claim 1. Given a TOM model with depth d, chunk size C and constant edge traversal probability p (with $0 < p < 1$). For a leaf w in the episodic tree,

$$P[w \text{ is recalled}] = p^d.$$

Due to space constraints, we only provide a quick intuition toward the proof of Claim 1. The idea is that a word w at depth d is recalled if and only if all edges on the path leading to w are traversed. This path is of length d, and each edge is traversed with probability p, implying the claim.

The following lemma addresses the main technical challenge underlying our analysis.

Lemma 1. *Consider a free recall test of k words and a simple-TOM model with depth d, chunk size C, with a subcritical semantic graph, and tree edge traversal probability p (with* $0 < p < 1$*). Then the expected number of recalled words is* $\Theta(k\, p^d)$.

Proof. Let $A \subseteq \{w_1, \ldots, w_k\}$ denote the set of recalled words using the episodic structure.

Let $C_1, \ldots, C_m$ denote the connected components of the semantic graph, where m is the number of connected components. Define the random variables

$$X_i = \begin{cases} |C_i|, & \text{if } C_i \cap A \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases}$$

Then, by a union bound,

$$\Pr[X_i = |C_i|] = \Pr[C_i \cap A \neq \emptyset] = \Pr\left[\bigcup_{c \in C_i} \{c \in A\}\right] \leq \sum_{c \in C_i} \Pr[c \in A] = |C_i| \cdot p^d.$$

Moreover, from Claim 1,

$$\Pr[X_i = |C_i|] \geq \max_{c \in C_i} \Pr[c \in A] = p^d$$

We now bound the expected number of recalled words, let R denote the number of recalled words. By linearity of expectation,

$$\mathbb{E}[R] = \mathbb{E}\left[\sum_{i=1}^{m} X_i\right] = \sum_{i=1}^{m} \mathbb{E}[X_i].$$

[4] With probability that tends to one as $k \to \infty$.

[5] For a graph G on a set of vertices V, the connected component of each vertex $v \in V$ is the set of all vertices reachable from v.

Using the upper bound on $\Pr[X_i = |C_i|]$, we obtain

$$\mathbb{E}[R] = \sum_{i=1}^{m} |C_i| \Pr[X_i = |C_i|] \leq \sum_{i=1}^{m} |C_i|^2 p^d = O\Big(k\, p^d\Big)$$

For the last transition, we used subcritical semantic graph definition. We now turn to obtain a lower bound. Similarly, using the lower bound on $\Pr[X_i = |C_i|]$,

$$\mathbb{E}[R] = \sum_{i=1}^{m} |C_i| \cdot \Pr[X_i = |C_i|] \geq \sum_{i=1}^{m} |C_i| \cdot p^d = k\, p^d,$$

which asymptotically matches the upper bound. □

Proof of Theorem 1. First, we claim that simple-TOM is monotonically increasing as a function of p, as the edge traversal probability increases, the probability of recalling words also increases. By monotonicity of simple-TOM we can define an interval $I = [p_{\min}, p_{\max}] \subseteq [0, 1]$ such that the expected number of recalled words is $\Theta\left(g\left(k\right)\right)$. $p_{\min}, p_{\max}$ would be the solution of the following equations:

- $k \cdot (p_{\min})^d = a \cdot g(k)$
- $k \cdot (p_{\max})^d = b \cdot g(k)$

The choice of constants $0 < a < b < k$ depends on how tight we want the expectation bounds to be. □

Simulations

We conducted comprehensive simulations of the Tree of Memory (TOM) model to evaluate its ability to capture key phenomena in free recall. The simulations focused on three aspects: single-list free recall behavior, the list-length effect under different attentional scaling regimes, and the interaction between episodic chunking and semantic retrieval.

Model Parameters

Unless otherwise noted, simulations use the following default parameter values. The episodic hierarchy employs a chunk capacity of $C = 4$ items per chunk. The attention function decays monotonically with position within each level according to $f(n) = 0.3 + 0.5^{n+1}$, where n denotes the zero-based position index.

In a similar spirit to (Katkov et al., 2022), the semantic component is modeled as an Erdős–Rényi (ER) random graph representing semantic associations. The semantic edge traversal probability is governed by the parameter $p_{\text{sem}}^{\text{edge}} = 0.1$.

All parameters are held fixed across experiments unless explicitly stated otherwise.

Experiment 1: Single-List Free Recall

To assess single-list free recall behavior, we simulated $N = 1000$ trials with lists of length $k = 20$. For each trial, a new episodic tree was constructed according to the hierarchical organization described in Section 2, and a new semantic graph was generated using the ER model with edge probability $q = 0.06$. Recall proceeded via the probabilistic depth-first search algorithm with semantic bursts, as described in Section 2.2.

From these simulations, we computed two standard measures of free recall performance:

- **Serial Position Curve (SPC)**: the probability of recalling an item as a function of its serial position.
- **Lag-Conditional Response Probability (lag-CRP)**: the probability of transitioning between recalled items as a function of their temporal lag.

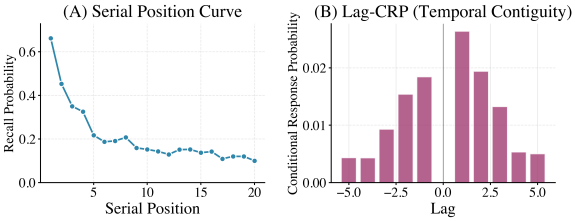

Figure 2: Single-list free recall simulation results. (A) shows the probability of recalling an item as a function of its serial position. (B) shows the probability of transitioning between recalled items as a function of their temporal lag.

Experiment 2: List-Length Effect

To examine the list-length effect, we simulated recall across a range of list lengths $k \in [10, 200]$ For each list length, we ran $N = 5000$ independent trials. We investigated two attentional scaling regimes that yield sublinear recall growth. In the first regime, corresponding to the logarithmic recall behavior predicted by the first bullet of Theorem 1, the attention function was set using

$$f(n) = \sqrt{3 \log(k) \cdot k^{-1}}. \tag{1}$$

This scaling follows from the theoretical bounds (taking lower bound with $a = 3$) on traversal probabilities and ensures logarithmic growth in expected recall. In the second regime, corresponding to the power-law behavior characterized in Theorem 1, we employed a weaker square-root scaling,

$$f(n) = \sqrt{2 \cdot k^{-1/2}}. \tag{2}$$

This choice permits faster recall growth while remaining strictly sublinear. For both regimes, the semantic graph edge existence probability was scaled inversely with list length, $q(k) = 0.5/k$, ensuring that the semantic component satisfied required conditions.

Expected recall performance was compared against theoretical predictions. Specifically, we considered $E[\text{recalls}] = \sqrt{3\pi k/2}$ for the power-law regime (Naim et al., 2020), and $E[\text{recalls}] = 3 \log k$ for the logarithmic regime.

Results

We now summarize the results of these simulations, focusing on how the TOM reproduces key empirical effects in free recall and how recall performance varies across parameter regimes.

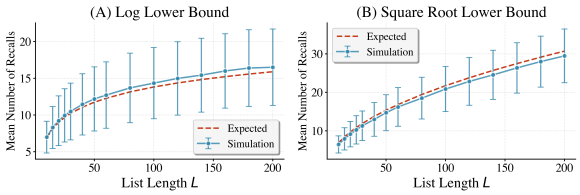

Figure 3: List-length effect simulations for (A) logarithmic and (B) power-law attention scaling.

Single-List Free Recall Figure 2(A) shows the serial position curve produced by the model. A pronounced primacy effect is observed for the first few list items, followed by uniformly low recall probabilities for middle positions. This pattern closely matches classic empirical findings (Kahana, 1996; B. B. Murdock, 1962). While TOM does not directly capture the recency effect, it can be naturally extended in a manner similar in spirit to SAM (J. Raaijmakers & Shiffrin, 1981) by incorporating a constant-capacity short-term store that retains the most recently studied words. At test, this store can support retrieval of the final list items. To more closely match the empirical recency effect, one may further assume that recall probability from the short-term store depends on recency, so that more recently presented words are more likely to be recalled than older items still maintained in short-term memory.

Figure 2(B) presents the lag-CRP, revealing elevated transition probabilities for small temporal lags and a clear forward asymmetry. This reflects the model's hierarchical episodic organization and the monotonic decay of the attention function, which jointly favor forward transitions between temporally adjacent items.

List-Length Effect Figure 3 shows recall performance as a function of list length for both attention-scaling regimes. In both cases, recall grows sublinearly with list length and closely follows the theoretical predictions.

The logarithmic regime produces slower growth, reflecting stronger attentional decay, whereas the power-law regime permits faster yet still sublinear recall growth. Together, these results capture the well-established empirical list-length effect.

Discussion

We introduced the *Tree of Memory* (TOM), a model in which free recall is generated by a probabilistic search over a hierarchical episodic representation, coupled to a minimal semantic association structure. The goal is not to replace fitted accounts of free recall (Howard & Kahana, 2002; Polyn et al., 2009; J. Raaijmakers & Shiffrin, 1981), but to test whether hierarchical episodic organization, instantiated at the level of list → chunk → item, can account for canonical free-recall benchmarks.

A central contribution is an explicit characterization of list-length scaling regimes within this hierarchical retrieval architecture. Episodic accessibility is governed by a monotone attention function over episodic edges, and recall corresponds to a probabilistic depth-first search (DFS). Under this setup, we derive regimes in which the expected number of recalled items grows either logarithmically or as a sublinear power law with list length, depending on how attention-weighted traversal probabilities scale. These results complement analytically characterized capacity laws derived in simplified settings (Katkov et al., 2022; Naim et al., 2020), while placing scaling within an explicit episodic hierarchy.

TOM also provides a mechanistic account of organizational effects. Primacy follows from the attention-weighted accessibility gradient, while temporal contiguity arises because episodic DFS tends to exhaust locally reachable subtrees (e.g., within chunks) before backtracking, promoting clustered transitions. The semantic component provides an additional retrieval route via associations triggered upon item recall. More broadly, this emphasis on hierarchical organization is consistent with evidence that imposed structure facilitates recall (Bower et al., 1969) and that hierarchical organization can emerge for random material (Naim et al., 2019; Zhong et al., 2026) and narrative (Zhong et al., 2025).

The current formulation is limited to single-list recall and does not address graded forgetting across lists or other multi-list effects. A natural extension is to attenuate episodic traversal probabilities with episodic remoteness (e.g., depth-dependent accessibility) while preserving the same hierarchical retrieval process. TOM also yields testable predictions: manipulations that steepen attentional decay across encoding should reduce both overall recall and temporal organization; manipulations that promote chunking should increase clustering and within-chunk transitions affecting performance (Cowan, 2001; Farrell, 2012); and increasing semantic association density (e.g., categorized lists) should amplify semantic-driven retrieval bursts, increasing recall without necessarily strengthening temporal contiguity (Bower et al., 1969).

Recent cognitive–AI agent-memory frameworks also motivate hierarchical/event-level representations paired with topology-aware retrieval (e.g., CompassMem (Hu et al., 2026), Amory (Zhou et al., 2026), GSW (Rajesh et al., 2026), STITCH (Yang et al., 2026)), where access is implemented as structured navigation over a memory graph rather than flat similarity search. These frameworks are complementary to TOM: they focus on scalable long-horizon systems, whereas TOM isolates a minimal hierarchical episodic representation with probabilistic DFS retrieval, yielding explicit predictions for recall organization and capacity scaling in single-list free recall.

Overall, TOM provides a constrained framework in which hierarchical episodic organization supports both recall organization and lawful capacity scaling, offering a basis for studying how attention-weighted access and semantic associations jointly shape free recall.

Acknowledgments

We made light use of AI-based tools during the development of this work. Cursor AI was used to assist with writing and debugging Python code for simulations of the TOM model. ChatGPT was used to assist with figure design, including model sketches, layout, coloring, and styling of visualizations. We stress that all scientific decisions, model design, analyses, and interpretations were performed by the authors.

Omri Ben-Eliezer is supported in part by an Alon Scholarship and by a Taub Family Foundation "Leaders in Science & Technology" Fellowship.

References

Bower, G. H., Clark, M. C., Lesgold, A. M., & Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorized word lists. *Journal of Verbal Learning and Verbal Behavior*, *8*(3), 323–343. https://doi.org/10.1016/S0022-5371(69)80124-6

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–185. https://doi.org/10.1017/S0140525X01003922

Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, *5*, 17–61.

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*(2), 223–271. https://doi.org/10.1037/a0027371

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

Hu, Y., Liu, J., Tan, J., Zhu, Y., & Dou, Z. (2026). Memory matters more: Event-centric memory as a logic map for agent searching and reasoning. https://doi.org/10.48550/arXiv.2601.04726

Janson, S., & Luczak, M. J. (2008). Susceptibility in subcritical random graphs. *Journal of Mathematical Physics*, *49*(12), 125207. https://doi.org/10.1063/1.2982848

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*(1), 103–109. https://doi.org/10.3758/BF03197276

Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychology*, *71*(Volume 71, 2020), 107–138. https://doi.org/https://doi.org/10.1146/annurev-psych-010418-103358

Katkov, M., Naim, M., Georgiou, A., & Tsodyks, M. (2022). Mathematical models of human memory. *Journal of Mathematical Physics*, *63*(7), 073303. https://doi.org/10.1063/5.0088823

Kupper, N., & Penrose, M. D. (2024). Largest component and sharpness in subcritical continuum percolation. https://api.semanticscholar.org/CorpusID:271212174

Murdock, B. (1960). The immediate retention of unrelated words. *Journal of Experimental Psychology*, *60*, 222–234. https://doi.org/10.1037/h0045145

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488. https://doi.org/10.1037/h0045106

Naim, M., Katkov, M., Recanatesi, S., & Tsodyks, M. (2019). Emergence of hierarchical organization in memory for random material. *Scientific Reports*, *9*, 10448. https://doi.org/10.1038/s41598-019-46908-z

Naim, M., Katkov, M., Romani, S., & Tsodyks, M. (2020). Fundamental law of memory recall. *Physical Review Letters*, *124*(1), 018101. https://doi.org/10.1103/PhysRevLett.124.018101

Penrose, M. (2003). *Random geometric graphs*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198506263.001.0001

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. https://doi.org/10.1037/a0014420

Raaijmakers, J., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93–134. https://doi.org/10.1037/0033-295X.88.2.93

Raaijmakers, J. G. W., & Shiffrin, R. M. (2002). Models of memory. In *Stevens' handbook of experimental psychology*. John Wiley Sons, Ltd. https://doi.org/https://doi.org/10.1002/0471214426.pas0202

Rajesh, S., Holur, P., Duan, C., Chong, D., & Roychowdhury, V. (2026). Beyond fact retrieval: Episodic memory for RAG with generative semantic workspaces. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*, 32782–32790. https://doi.org/10.1609/aaai.v40i39.40557

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*(1), 156–163. https://doi.org/10.1016/S0022-5371(67)80067-7

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/BF03209391

Yang, R., Jiang, Y., Jiang, Y., Kargupta, P., Zhang, Y., & Han, J. (2026). Grounding agent memory in contextual intent. https://doi.org/10.48550/arXiv.2601.10702

Zhong, W., Can, T., Georgiou, A., Shnayderman, I., Katkov, M., & Tsodyks, M. (2025). Random tree model of meaningful memory. *Physical Review Letters*, *134*(23), 237402. https://doi.org/10.1103/g1cz-wk11

Zhong, W., Katkov, M., & Tsodyks, M. (2026). Synaptic theory of chunking in working memory. *eLife*, *15*. https://doi.org/10.7554/eLife.109538

Zhou, Y., Guo, X., Bayar, B., & Sengamedu, S. H. (2026). Amory: Building coherent narrative-driven agent memory through agentic reasoning. *Proceed-*

ings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 3926–3938. https://doi.org/10.18653/v1/2026.eacl-long.183