

Bridging Real-world Summarization and Laboratory-based Memory Recall

Hemali Angne ^a, Snigdha S. Mishra ^a, Guo Lin ^a, and Qiong Zhang ^{a,b,c,*}

^aDepartment of Computer Science

^bDepartment of Psychology

^cRutgers Center for Cognitive Science

Rutgers University–New Brunswick

Piscataway

NJ

USA 08854

*Correspondence to: qiong.z@rutgers.edu

Author Note

Word count for the text excluding title, references, author affiliations, acknowledgments, figures, and figure legends is 8,373. A subset of the results in Analysis 1 has appeared in a conference article in the non-archival conference proceedings of the Annual Meeting of the Cognitive Science Society in July 2025. The analysis code and datasets used in the study can be accessed at: <https://osf.io/bx2n6> (Angne, 2026). H.A. contributed to: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing – original draft, and writing – review & editing. S.S.M. and G. L. contributed to: data curation, formal analysis, and software. Q.Z. contributed to: conceptualization, investigation, methodology, supervision, writing – original draft, and writing – review & editing.

Abstract

A fundamental question in psychology is how humans digest, retain, and reconstruct information. While this process has been extensively studied in laboratory-based memory recall tasks, we have missed opportunities to examine it under more naturalistic, real-world conditions. Summarization is a task that people routinely engage in every day, and it is also a long-standing challenge in the machine learning community. Rich datasets of human summaries from machine learning provide a valuable, untapped resource for cognitive science researchers. In this work, we analyze 45,220 human summaries drawn from three large datasets, including news articles, autobiographical stories, and Reddit posts, and compare them with a dataset of narrative memory recall. We demonstrate a close relationship between human summarization and laboratory-based memory recall. While summarization operates at a higher level of abstraction than memory recall, the two tasks share deep similarities. The process of summarizing is fundamentally constrained by memory mechanisms, displaying well-known memory regularities, including primacy, recency, and temporal contiguity effects. Conversely, memory recall takes the form of summarization, prioritizing central and important information. Together, these findings bridge the real-world task of summarization and the rich literature on memory recall and lay the foundation for developing integrated theories of how humans digest and reconstruct information.

Public Significance Statement: A central question in psychology is how we encode and reconstruct experiences. While this is well-studied in laboratory memory recall tasks, less is known about how it functions in everyday contexts. In this work, we focus on summarization, a task people routinely engage in. We analyze over 45,000 human-written summaries from datasets developed in machine learning and find connections between how people summarize and remember. We find that summarization reflects characteristics of memory recall, as we can only summarize from what we remember. Additionally, our memories aren't perfect recordings; they're like summaries, focusing on the more important information. These findings facilitate cross-domain knowledge exchange, providing a different way to understand how our brains learn and reconstruct information in naturalistic contexts.

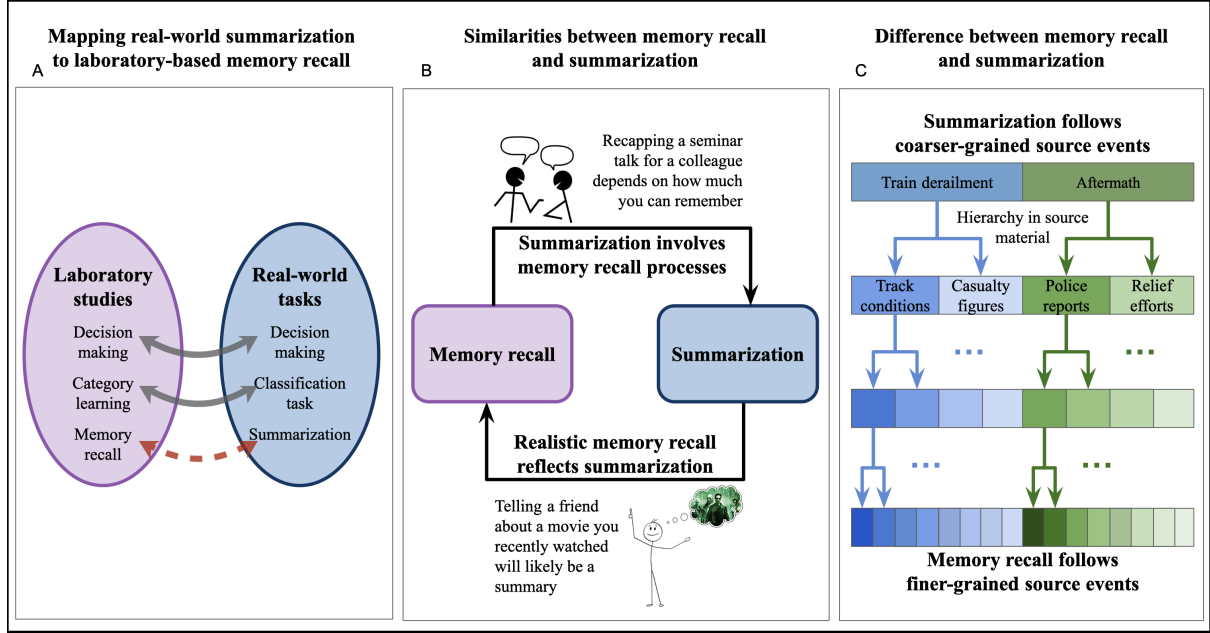
Keywords: summarization, memory recall, naturalistic conditions, natural language processing, network analysis

Bridging Real-world Summarization and Laboratory-based Memory Recall

Introduction

A central question in psychology is how we encode and reconstruct experiences. While this is well-studied in laboratory memory recall tasks (Kahana, 2012), we have missed opportunities to examine it under more naturalistic, real-world conditions. Research across many subdomains of human cognition has increasingly relied on rich, real-world datasets to generalize existing findings and reveal new insights. For instance, large-scale analyses of human planning have utilized response times across 12 million chess games (Russek et al., 2025), and studies on the transfer of learning have examined online cognitive training data from over 36,000 users (Steyvers & Schafer, 2020). Studying how we encode and reconstruct experiences, however, presents a unique challenge: the data analyzed must contain both the original experience and the reconstructed experience. Real-world data sources rarely provide both. Platforms like Google Memories record the objective original experience through photos and videos, but lack the reconstruction; platforms like Airbnb capture reconstructed experiences through user reviews, but lack the ground truth of the original event. To overcome this limitation, we propose utilizing the text summarization task. We constantly engage in summarization – such as recapping a seminar, summarizing a meeting, or taking notes on a paper. Summarization is also a long-standing problem in natural language processing (Erkan & Radev, 2004; T. Zhang et al., 2024), where large datasets of human summaries are curated as references for training machine learning models (for a review, see H. Zhang et al., 2025). Despite its ubiquity in our daily lives and its importance for applications in machine learning, the link between laboratory-based memory recall and real-world summarization remains largely unexplored (Figure 1A). To bridge this gap, we propose a mapping between the two tasks, hypothesizing a close relationship between them. The advantage of this approach is that we can transfer established methods from the vast literature on human memory, leveraging over a century of accumulated knowledge to understand summarization. Simultaneously, we can repurpose summarization datasets from machine learning to understand human memory in naturalistic conditions.

We formally analyze the relationship between summarization and recall, including both the similarities and differences between them. We first hypothesize that summarization involves fundamental processes of memory recall. Since the objective of summarization is to

**Figure 1**

Our proposed framework mapping summarization to memory recall. (A) A foundational approach in cognitive psychology is to deconstruct complex real-world behaviors into isolated, well-controlled laboratory tasks. In this work, we study the relationship between laboratory-based memory recall and real-world summarization. (B) We propose a bi-directional relationship between the summarization task and the memory recall task, where summarization involves processes of recall and recall reflects characteristics of summarization. (C) There are also differences between the two tasks, as we hypothesize that summarization operates on more coarse-grained levels of event hierarchy, whereas memory recall focuses on more fine-grained levels.

retain key information from the source material, before such information is condensed or summarized, we should consider summarization first as a task of memory recall. Consider a real-world summarization example: when recapping a seminar talk for a colleague (Figure 1B, top), any details forgotten from the talk will not make it into the summary, regardless of their importance. In other words, the ability to generate summaries is constrained by the ability to recall information from the original source. If this hypothesis holds, we should expect to see alignment between patterns in human summaries and known patterns in human memory recall. Specifically, when subjects are asked to recall in any order all the information they can remember from a list (as studied in a free recall task; Murdock, 1962; Roberts, 1972), there are typical patterns associated with the temporal dynamics of their memory recalls, such as where

one’s recall is likely to start from (relative to the original material) and where one’s next recall is likely to transit to (Kahana, 1996; Murdock, 1962). Inspired by this literature in examining the temporal dynamics of memory recalls, we will carry out the same analyses to understand summarization by examining where in the source material a summary tends to begin and how it unfolds from one sentence to the next, looking for signatures of classic memory effects.

In addition to showing that summarization involves processes of memory recall, we hypothesize that memory recall reflects processes of summarization. This is a relationship that can be overlooked in well-controlled laboratory experiments that demand verbatim recall. In more naturalistic settings, however, when we have limited time and effort, recall often takes the form of summarization. For instance, when describing a recently watched movie to a friend (Figure 1B, bottom), we would naturally provide a summary of it rather than going over every single detail. In other words, during naturalistic memory recall, we tend to retrieve the most important information and leave out any peripheral details, which is the essence of summarization. While previous work has supported that recall undergoes temporal compression compared to the original experience (Georgiou et al., 2025; Jeunehomme & D’Argembeau, 2020; Musz & Chen, 2022), it leaves open the question of exactly what form this compression takes, and whether it resembles that of a summarization task. To answer this question, we will examine whether naturalistic memory recall shows patterns typically observed in summarization studies. One such pattern is the effect of centrality, where information important and central to the original source has a higher likelihood of being included in a summary (Brown et al., 1983; Omanson, 1982; Winograd, 1984). Although prior memory literature has supported the role of an event’s semantic centrality in predicting its recall probability (Antony et al., 2024; Lee & Chen, 2022), here we conduct a direct comparison between summarization and recall studies over these analyses, providing an opportunity to reinterpret these memory effects based on their similarity with summarization.

While summarization and recall share similarities, we hypothesize that there are also important differences between them. Although both tasks involve reconstructing information from the original source, we hypothesize that they function at different hierarchies of the source. Prior research has shown that humans encode experiences by segmenting them into meaningful events organized across multiple hierarchical levels (Baldassano et al., 2018). We propose that summarization operates on more coarse-grained, high-level events, whereas

memory recall focuses on more fine-grained ones. For example, a news article about a train derailment might contain the coarse-grained event ‘train derailment’ (Figure 1C, in blue), which further encompasses fine-grained events like track conditions and casualty figures. This could be followed by a second coarse-grained event ‘aftermath’ (Figure 1C, in green), which includes fine-grained events of police reports and relief efforts. A summary would likely follow the high-level progression from ‘train derailment’ to ‘aftermath’, while a detailed recall would navigate the finer-grained events within each. To test this hypothesis, we segment each source into events at multiple hierarchies, as illustrated in Figure 1C, and analyze how summaries and recalls transition between these events.

We carry out the outlined analyses on three publicly available summarization datasets, containing 302 summaries of news articles (T. Zhang et al., 2024), 2,779 summaries of autobiographical events (The HIPPOCORPUS dataset; Sap et al., 2022), and 42,139 summaries of Reddit posts (Kim et al., 2019). These datasets of human summaries were previously developed in the machine learning community for evaluating tools of natural language processing and models of automated summarization. To ensure that the similarities or differences we see between the summarization datasets and the free recall task are not a result of stimuli richness (summarization tasks typically use continuous texts while standard free recall tasks use discrete word lists), we additionally analyzed a dataset of narrative free recall (Georgiou et al., 2025), which is a naturalistic version of the free recall task, involving continuous, semantically related source material. To foreshadow our results, the analysis of temporal dynamics reveals that there is an alignment in summarization patterns and memory recall patterns, suggesting that memory constrains the way we summarize. Furthermore, we found that naturalistic memory recall exhibits a centrality effect, where important information is more likely to be recalled, mirroring a key feature of summarization. Finally, the analysis of transition patterns at multiple source hierarchies shows where the two tasks differ, with memory recall following more fine-grained source events, while summarization operating at a coarser level. Together, these results provide us with a better understanding of the cognitive processes involved in summarization. Linking the summarization task with the memory recall task also opens up future opportunities to study each task, informed by the knowledge and datasets from the other.

Results

Analysis 1: Summarization involves processes of memory recall

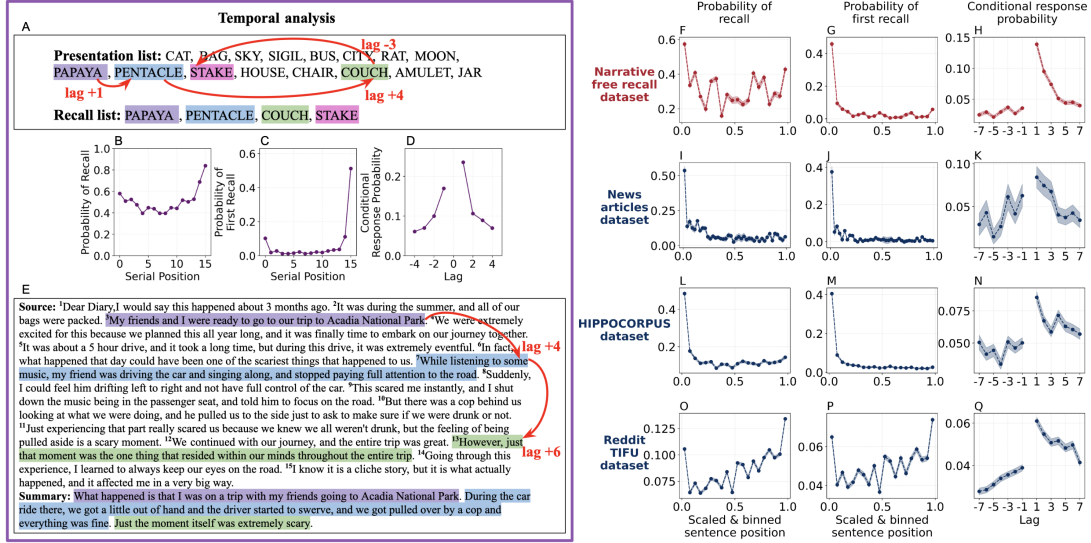


Figure 2

Temporal dynamics in summarization align with patterns observed in memory recall. (A) The standard free recall task is illustrated, where the goal is to recall as many items as possible from the presentation list in any order. Items recalled successively correspond to items presented either nearby (e.g., lag +1) or far away (e.g., lag -3, lag +4). Averaging across multiple trials gives behavioral patterns in standard free recall relative to the presentation list: (B) the serial position curve, showing which item positions are more likely to be recalled; (C) the probability of first recall, showing which item position recall is likely to start from; and (D) the conditional response probability, indicating how likely it is to successively recall items studied at nearby positions. These patterns are plotted from the free recall stage of the study in Cornell et al. (2024). (E) A trial of the summarization task from the HIPPOCORPUS dataset is shown, where the goal is to retain key information from the source material. We identified, for each sentence in the summary, which sentence in the source it best mapped to based on a similarity metric, the METEOR score. Successive sentences in a summary correspond to either nearby (e.g., lag +4) or far away (e.g., lag +6) sentences in the source material. The same behavioral patterns as in standard free recall were analyzed for the (F–H) Narrative free recall, (I–K) News articles, (L–N) HIPPOCORPUS, and (O–Q) Reddit TIFU datasets. The shaded error represents the standard error of the mean.

Examining alignment between summarization and memory recall in their temporal dynamics. We first tested whether summarization involves processes of memory recall. Specifically, we analyzed the narrative free recall and summarization datasets for effects typically seen in standard free recall. The standard free recall task, which has been studied for decades to gain insight into the mechanisms and representations underlying memory search (Murdock, 1960, 1962; Roberts, 1972; Standing, 1973), involves subjects studying a list of items (such as words) and then recalling as many items as they can in any order (as seen in the illustrative trial in Figure 2A). Averaging across trials, temporal dynamics in recall behavior can be analyzed, like which item positions are more likely to be recalled (the serial position curve) and where recall is likely to start relative to the presentation list (the probability of first recall curve). These patterns are plotted in Figures 2B–2C from a free recall study in Cornell et al. (2024). The serial position curve demonstrates a general tendency of subjects to have an enhanced recall for items from the beginning or the end of the list, while the probability of first recall curve demonstrates a similar tendency in where recall is initiated (primacy or recency effects; Murdock, 1962). Both primacy and recency effects are robust across multiple studies, though their relative ratio can vary given task conditions (Howard & Kahana, 1999; Tan et al., 2016) and how well subjects perform the task (Q. Zhang et al., 2023). In addition to recall probability and initiation, one can also examine recall transitions by analyzing how likely it is to successively recall items studied at nearby serial positions in the presentation list (conditional response probability, computed by dividing the number of observed transitions for a lag by the number of possible transitions for that lag; Kahana, 1996). The conditional response probability demonstrates a greater tendency for successive recalls to be from nearby serial positions, with smaller lags, than farther away positions with larger lags (temporal contiguity effects; Kahana, 1996). For example, Figure 2A has a nearby transition of lag +1 going from the first recall ‘PAPAYA’ (at serial position 9) to the second recall ‘PENTACLE’ (at serial position 10), and a far away transition of lag +4 going from ‘PENTACLE’ to ‘COUCH’. Subjects make both backward transitions (at negative lags) and forward transitions (at positive lags), though with a greater tendency for forward transitions (forward asymmetry; Kahana, 1996). Temporal contiguity and forward asymmetry effects can be seen in Figure 2D plotted from the same study (Cornell et al., 2024). These well-documented patterns in standard free recall provided a framework for analyzing similar

dynamics in narrative free recall and summarization data. To apply this framework, sentences in the source material can be considered analogous to items in a presentation list, while sentences in the summary (or recall) are analogous to recalled items. Following this analogy, we identified which sentence in the source each summary sentence best mapped to, based on maximum semantic alignment using METEOR score (Banerjee & Lavie, 2005). Using these mappings, we can obtain the serial position curve by estimating which sentences in the source material are more likely to be included in the summary, and the probability of first recall curve by estimating where the summary is likely to start from relative to the source material. We normalized sentence positions in the source material to a range of 0 to 1 and used fixed-width bins to average across trials, enabling us to plot source material of varying lengths together. Because primacy and recency effects in standard free recall vary based on task conditions and subject pool, we expect these temporal patterns to be dataset-dependent for naturalistic stimuli as well. Lastly, we can similarly derive the conditional response probability to examine transition patterns (in lags) by tracking the positions in the source material that successive summary sentences map to. As shown in the example in Figure 2E, the first and second summary sentences are mapped to position 3 and position 7 in the source material, respectively, resulting in a lag of +4 between them.

Summarization datasets demonstrate memory regularities similar to those observed in standard free recall literature. To test whether the memory biases of primacy and recency were significant, we compared the likelihood of either recalling or initiating a recall from the first or last source sentence, respectively, against the middle three source sentences. In the Narrative free recall, News articles and HIPPOCORPUS datasets, subjects were more likely to recall or summarize sentences from the beginning of the source material, showing a primacy effect ($p < .001$ for all comparisons, determined by non-parametric permutation tests in which the position labels of binary recall outcomes were randomly shuffled 1,000 times separately for each dataset; Figures 2FIL). In the Narrative free recall, HIPPOCORPUS, and Reddit TIFU datasets, they were also more likely to include sentences from the end, showing a recency effect ($p < .001$ for all comparisons, determined by the same permutation tests; Figures 2FLO). These patterns in the datasets aligned with those typically observed in standard free recall (as seen in Figure 2B). The probability of first recall curves for the Narrative free recall, News articles and HIPPOCORPUS datasets (Figures 2GJM)

showed a primacy effect, with subjects showing a tendency to initiate their recall or summary from the beginning of the source text ($p < .001$ for all comparisons, determined by the same permutation tests). Furthermore, the Narrative free recall and Reddit TIFU datasets exhibited a recency effect, with subjects more likely to start their recall from the end of the text ($p = .009$ and $p = .001$, respectively, determined by the same permutation tests; Figure 2GP). Both of these patterns are typical of standard free recall tasks (Figure 2C). As a summary (or recall) continued to unfold, it was more likely for successive summary sentences to be from nearby sentence positions in the source (lags ± 1) than farther away ones (lags ± 2 through ± 7) in the Narrative free recall ($p < .001$), News articles ($p = .017$), HIPPOCORPUS ($p < .001$), and Reddit TIFU ($p = .026$) datasets (determined by non-parametric permutation tests in which recall sequences were randomly shuffled 1,000 times; Figures 2HKNQ). Moreover, there was a greater tendency for forward transitions at positive lags (+1, +2, +3) than backward ones at negative lags (-1, -2, -3) across all datasets ($p < .001$ for all comparisons; determined by the same permutation tests). Both these transition effects aligned with the temporal contiguity and forward asymmetry effects typically observed in standard free recall (Figure 2D). These conclusions hold when using alternative methods of mapping summary sentences to source sentences (see section S1. in the Supplementary Materials), as well as when examining two additional naturalistic recall datasets (see section S2. in the Supplementary Materials). The observation that the summarization datasets demonstrated primacy, recency, temporal contiguity, and forward asymmetry effects, similar to those observed in standard free recall literature, supports our hypothesis that summarization involves processes of memory recall. The similarity between behavioral patterns in the Narrative free recall dataset (Figures 2F-2H) and the three summarization datasets (Figures 2I-2Q) provides further evidence for this hypothesis. The content of a summary is constrained by what can be retrieved from memory. People tend to recall the beginning and end of a source better (primacy and recency effects), so these parts are more likely to be included in a summary. Likewise, sentences that are close together in the source material are often recalled together and, in turn, placed together in a summary. These findings suggest that memory constraints and mechanisms shape how people summarize information.

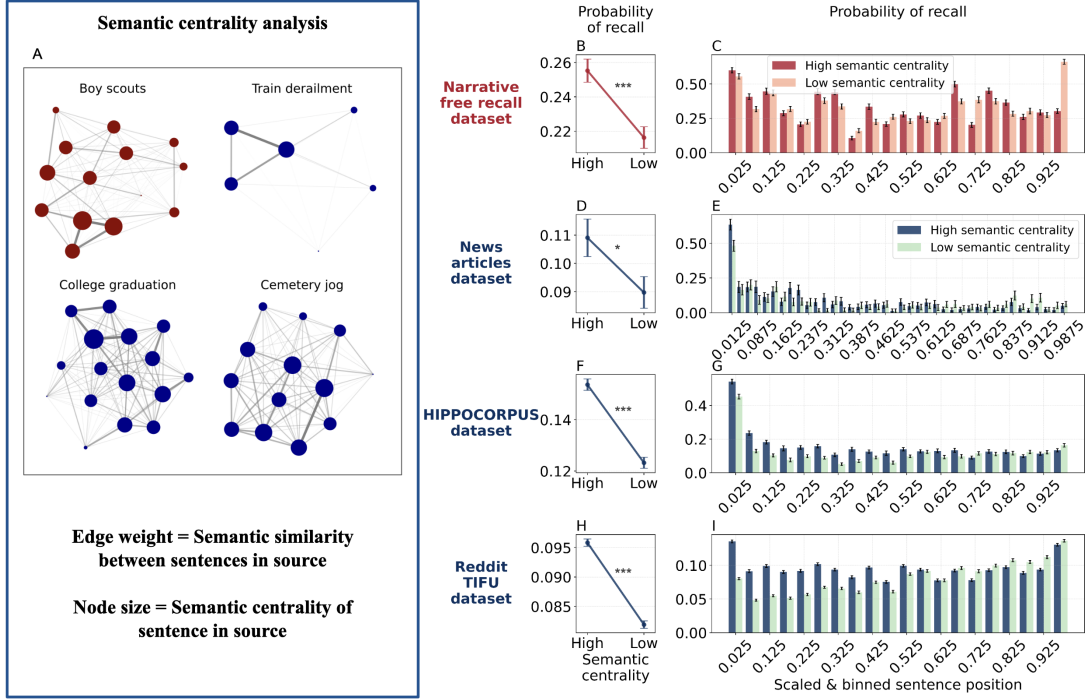


Figure 3

More important and central source elements are more likely to be recalled, similar to the patterns observed in summarization. (A) For each source, we constructed a graph in which sentences are represented as nodes and edge weights represent semantic similarity between sentence pairs. The semantic centrality of a sentence node was computed using the PageRank algorithm, which estimates a sentence's importance based on its connections to other sentences and their importance. Source sentences were then assigned to high (top 40% by centrality) and low (bottom 40%) semantic centrality groups. In the (B–C) Narrative free recall, (D–E) News articles, (F–G) HIPPOCORPUS, and (H–I) Reddit TIFU datasets, the probability of recall was greater for the sentences in the high semantic centrality group than in the low semantic centrality group. This centrality advantage was primarily centered around sentences appearing early in the source. The error bars represent the standard error of the mean.

Analysis 2: Memory recall reflects characteristics of summarization

Examining the alignment between memory recall and summarization based on the centrality effect. Next, we tested whether naturalistic memory recall exhibits the characteristics of summarization. A well-documented pattern in summarization is the centrality effect – the tendency of subjects to identify and include elements central to the source in their summaries (Brown et al., 1983; Denhiere et al., 2005; Omanson, 1982;

Winograd, 1984). While there are several ways of quantifying centrality, we used an automated approach from prior work that does not rely on human annotations. In this approach, a sentence’s centrality is defined as the strength of connections between that sentence and every other sentence in the source (Denhiere et al., 2005; Kintsch, 2008). Our analysis focused specifically on centrality defined by semantic relations. We constructed a graph for each source by representing sentences as nodes and assigning edge weights based on the semantic similarity between each pair of sentences, measured using the METEOR score (Banerjee & Lavie, 2005). Higher similarity corresponds to heavier edge weights as visualized in the example trials in Figure 3A. Within this graph representation, we computed sentence centrality using the PageRank algorithm that estimates how connected one sentence is to other sentences (Griffiths et al., 2007; Page et al., 1999). As illustrated in Figure 3A, more semantically central sentences received higher PageRank scores (visualized as larger nodes). We then assigned sentences to high (top 40% by centrality) and low (bottom 40%) semantic centrality groups within each trial and calculated the probability of recall for each group. Averaging across trials, we tested for the centrality effect – whether sentences in the high semantic centrality group are more likely to be included in recalls (or summaries).

The Narrative Free Recall dataset demonstrates higher recall for sentences most central to the source, similar to patterns seen in summarization. We found that sentences semantically central to the source were more likely to be included in recalls (or summaries) across all datasets. Specifically, the probability of recall was greater for the high semantic centrality group than the low semantic centrality group (Figures 3BDFH) and this difference was significant for the Narrative free recall (Wilcoxon signed-rank test:

$W = 75661.0, n = 843, p < .001$, two-tailed), News articles (Wilcoxon signed-rank test:

$W = 9182.5, n = 302, p = .018$, two-tailed), HIPPOCORPUS (Wilcoxon signed-rank test:

$W = 776654.5, n = 2779, p < .001$, two-tailed), and Reddit TIFU (Wilcoxon signed-rank test:

$W = 245652895.5, n = 42036, p < .001$, two-tailed) datasets. Building on this overall centrality

effect, we next examined its interaction with sentence positions within the source. For each bin position, we grouped the sentences by high (top 40% by centrality) and low (bottom 40%) semantic centrality across trials, and calculated the probability of recall associated with each group. We fit a logistic regression model to examine the effects of semantic centrality group and bin position on the likelihood of a sentence being included in a recall (or summary).

Across all datasets, sentences in the high semantic centrality group were significantly more likely to be recalled than those in the low semantic centrality group (Narrative free recall $OR = 1.35, 95\%CI [1.21, 1.51], p < .001$; News articles $OR = 1.73, 95\%CI [1.32, 2.26], p < .001$; HIPPOCORPUS $OR = 1.88, 95\%CI [1.69, 2.09], p < .001$; Reddit TIFU $OR = 1.96, 95\%CI [1.90, 2.03], p < .001$; Figures 3CEGI). However, this advantage diminished for sentences appearing later in the source, as indicated by a significant negative interaction between semantic centrality group and bin position (Narrative free recall $OR = 0.97, 95\%CI [0.96, 0.98], p < .001$; News articles $OR = 0.97, 95\%CI [0.95, 0.98], p < .001$; HIPPOCORPUS $OR = 0.96, 95\%CI [0.95, 0.97], p < .001$; Reddit TIFU $OR = 0.95, 95\%CI [0.95, 0.96], p < .001$; Figures 3CEGI). The centrality advantage is strongest for early source sentences, likely because they have more opportunities to be reactivated. High centrality sentences that appear early share semantic content with several later sentences. As subjects continue reading, these later sentences repeatedly reactivate earlier ones, strengthening their representation in memory and increasing their likelihood of being recalled. These conclusions also hold when examining the same centrality effect across two additional naturalistic recall datasets (see section S2. in the Supplementary Materials). Together, these findings support our hypothesis that naturalistic memory recall exhibits characteristics of summarization, as the Narrative free recall dataset demonstrated centrality effects similar to those observed in the summarization datasets. Although accurately recalling all the information from the source is the explicit goal of a memory task, people do not spend infinite effort and time searching their memory. Under limited time and cognitive resources, more important and central information is prioritized during recall, which is the essence of summarization.

Analysis 3: Summarization and recall operate at different levels of hierarchy in the source

Examining transition patterns in datasets at multiple hierarchies of the source. We next analyzed whether summarization and memory recall differ in their transition patterns when analyzed at multiple hierarchies of the source. Given that summarization explicitly requires condensation of the source material, we tested the hypothesis that summarization operates at more coarse-grained levels of the source, whereas memory recall operates at more fine-grained levels. Using a method for predicting event boundaries from

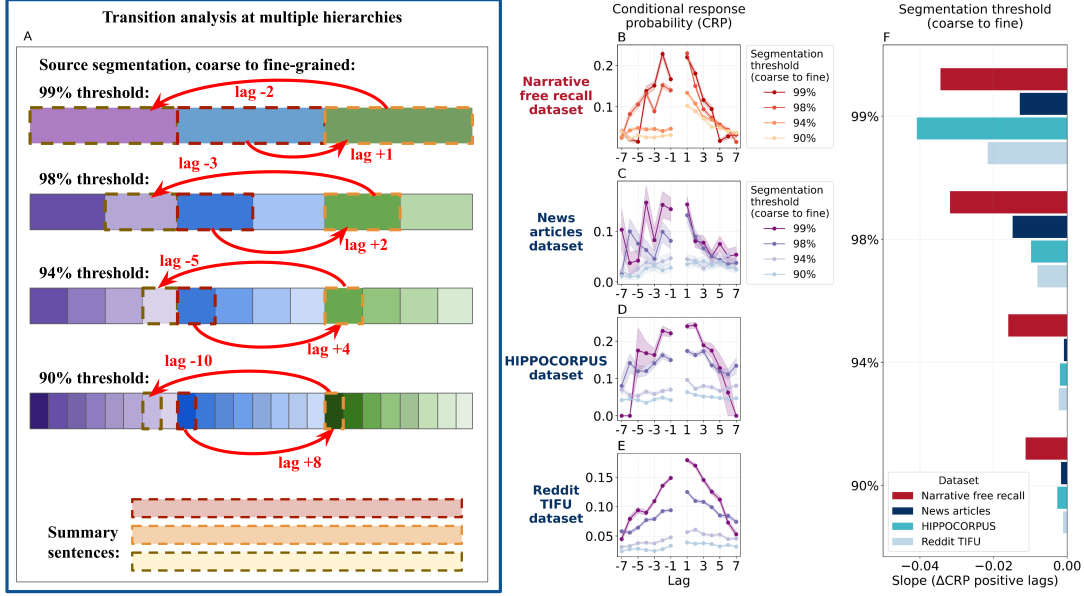


Figure 4

Summarization and recall make temporal transitions at different levels of event hierarchy. (A) Each source was segmented at four different thresholds to get progressively finer events. Summary sentences were mapped to these source events based on the similarity metric of METEOR score, and transition patterns (in lags) were examined by tracking the event positions to which successive summary sentences map. Conditional response probability curves at four segmentation thresholds were plotted for (B) the Narrative free recall dataset and (C–E) the summarization datasets. For each dataset, we fit a linear regression model to predict conditional response probability from lags (+1 to +7) across the four segmentation thresholds. (F) The Narrative free recall dataset exhibited steep negative slopes (indicating a strong temporal contiguity effect) at all segmentation thresholds, whereas summarization datasets showed flatter slopes at finer thresholds of 94% and 90%. The shaded error represents the standard error of the mean.

Kumar et al. (2023), we segmented the source into events at four segmentation thresholds. The segmentation threshold decides how coarse-grained or fine-grained the event segmentation will be. We used thresholds of 99%, 98%, 94%, and 90% to get progressively finer source events as illustrated in Figure 4A. For each segmentation threshold, we tokenized each summary (or recall) sentence and all source events. Each summary sentence was then mapped to a source event based on the maximum similarity metric of METEOR score (see Figure 4A, where mappings are indicated by dashed-line boxes of corresponding colors). We then computed the

conditional response probability to analyze the likelihood of successive summary (or recall) sentences to make nearby transitions as opposed to distant ones between source events.

The Narrative free recall dataset maintains temporal contiguity at all levels of event hierarchy while summarization datasets operate at more coarse-grained levels. Figures 4B–4E show that at more coarse-grained levels (thresholds of 99%, and 98%), both recalls and summaries were more likely to transition between nearby source events at smaller lags than to far away ones at larger lags, showing a strong temporal contiguity effect. However, at more fine-grained levels (thresholds of 94%, and 90%), only the Narrative free recall dataset continued to exhibit strong temporal contiguity for positive lags (Figure 4B), while the curves for the summarization datasets flattened (Figures 4C–4E), indicating that summaries make fewer nearby transitions in the source at these levels. To quantify the strength of temporal contiguity effects across different thresholds in each dataset, we fit linear regression models at each segmentation threshold, predicting conditional response probability as a function of positive lags (+1 to +7). Figure 4F shows that both the Narrative free recall and summarization datasets exhibited steep negative slopes at the coarsest segmentation thresholds of 99% and 98%, showing considerable temporal contiguity. However, at finer thresholds of 94% and 90%, only the Narrative free recall dataset continued to show steep negative slopes, while the summarization datasets showed much flatter ones. The slope differences between the narrative free recall data and the summarization datasets were significant at these finer thresholds, as the probability of slope differences as large as those observed between narrative recall and each summarization dataset occurring by chance was low ($p < .001$ for all comparisons; determined by non-parametric permutation tests in which trial-level dataset labels were randomly shuffled 1,000 times separately for each Narrative free recall and summarization dataset pairing, and for each of the two thresholds). We additionally identified that the observed negative slopes at finer thresholds in the narrative free recall data are primarily driven by recall of short source narratives, whereas the recall of longer source narratives demonstrates similar flattened slopes as those in the summarization datasets (see section S2. in the Supplementary Materials). This hypothesis regarding the role of narrative source length in temporal transition patterns is further validated in two additional naturalistic recall datasets we analyzed, with source lengths ranging from 5 to 7 sentences in Martinez (2024) and from 72 to 190 sentences in Racciah et al. (2024) (see section S2. in the

Supplementary Materials). This dependence on source length, in addition to the task objective (summarization or recall), is consistent with a previous study showing for the longest narratives, some subjects tend to compress several clauses in the original narrative into a single clause of recall, as reported in the Supplementary Materials of Georgiou et al. (2025). Together, these findings demonstrate where memory recall and summarization differ. Memory recall aims to retrieve as much information as possible from the source, while summarization involves retaining and conveying the most central information. Consequently, summarization omits finer details and transitions only between more coarse-grained source events; in contrast, recall transitions between events at all levels with shorter source lengths, but increasingly resembles summarization as source length and memory demands increase.

Discussion

In this work, we aim to bridge the study of summarization and the literature of memory recall. Both tasks involve reconstructing information from a source, but with different primary objectives: summarization aims to distill the most important information, while laboratory recall tasks emphasize recovering as much verbatim information as possible. This leads to different temporal transition patterns; we demonstrate that summaries operate on a higher level of the event hierarchy, whereas memory recall operates at a more fine-grained level and is additionally regulated by source length. However, our most compelling findings highlight the shared mechanisms between these two processes. The distinction between recovering the gist for summarization and verbatim details for recall is not as clear-cut as assumed. We show that summarization exhibits primacy, recency, and temporal contiguity effects – hallmarks of behavioral patterns during free recall. This suggests that the process of summarizing is fundamentally constrained by the mechanisms of memory recall. Additionally, we demonstrate that recall is not merely verbatim retrieval; it is also guided by the importance of the material under cognitive constraints. This centrality effect, where more central and important information is prioritized, appears in both our summarization and narrative recall data. Taken together, these findings challenge the traditional separate treatment of the two tasks and point towards shared cognitive mechanisms. We now turn to the broader implications of these results for understanding both summarization and memory recall.

The literature studying summarization has primarily focused on examining the factors that relate to one’s ability to provide good summaries. For example, summarization behavior and ability have been associated with different developmental ages (Brown et al., 1983; Winograd, 1984), comprehension ability (McNamara et al., 2024; Ophuis-Cox et al., 2024), attention (Taieb-Maimon et al., 2024), and working memory capacity (Kintsch & Van Dijk, 1978). While early theoretical work outlined how summaries are generated based on principles of text comprehension (Kintsch & Van Dijk, 1978; Rumelhart, 1977) and identification of central units (Lehnert, 2014), they are not widely integrated with more recent empirical findings. One factor that may have hindered the speed of knowledge accumulation in the literature of summarization is the naturalistic nature of its stimuli and paradigms, making it difficult to unify empirical results conducted under different experimental settings. In other sub-fields, cognitive psychology addresses this challenge by creating a small set of simplified,

well-controlled laboratory paradigms (Mandler, 2011). Our understanding of memory, decision-making, and categorization, for instance, is built upon such standardized tasks, whose simplicity and tractability facilitate theory development. However, this approach is fundamentally incompatible with summarization. To simplify the task by removing the rich, continuous, and diverse source materials would be to remove the very essence of what is being studied in summarization.

To circumvent this challenge, in this work, we link summarization to an existing well-studied laboratory task, the free recall task, and compare summarization with a naturalistic version of the memory task. Recent work in cognitive psychology has increasingly realized the importance of generalizing and testing theory directly in more naturalistic settings and developed computational methods in doing so (Carvalho & Lampinen, 2025; Q. Zhang, 2026). By analyzing the commonalities and differences between naturalistic memory recall and summarization, we could transfer knowledge from memory literature to understand the cognitive processes of summarization, and from the task goal of summarization to better interpret known memory regularities. Our central finding is that the temporal dynamics of summarization align with patterns seen in free recall, including primacy, recency, and temporal contiguity effects. These patterns are not just empirical regularities in memory, but have contributed to constraining and developing theories of memory on how information is encoded and later retrieved. These patterns are explained by the Temporal Context Model (TCM), which posits that one’s current context drifts slowly toward each new experience during study and is later used to drive recall (Howard & Kahana, 2002; Polyn et al., 2009). Recency effects arise because the context at the start of recall is close in time to the context at the end of study, which is associated with end-of-list items. Temporal contiguity effects occur because items studied nearby in the list are tied to similar context states during encoding, so that when an item is recalled, it reinstates its original study context and subsequently activates its surrounding items. Primacy effects occur as items at the beginning of the list receive more attention. Linking patterns in summarization to known patterns in free recall allows us to transfer knowledge from memory literature and provide a mechanistic explanation for why summarization prioritizes information from the beginning and end of a source text, and why transitions often occur between temporally adjacent information.

Aligning the tasks of summarization and recall also deepens our understanding of

known regularities in memory literature. We observed a centrality effect in both memory recall and summarization, where more important and central source elements were more likely to be recalled or summarized. This effect in memory recall has been documented previously (Antony et al., 2024; Lee & Chen, 2022), as an event’s semantic centrality – the sum of its weighted connections to other events – has been shown to positively predict its likelihood of being recalled. This effect was attributed to inter-event connections benefiting memory both at encoding and at recall. During encoding, highly interconnected events are more likely to be reactivated by each other, strengthening their representation in memory. During recall, high centrality events have stronger associative links to other events, making them more likely to be cued (Lee & Chen, 2022). While this explanation provides a mechanistic account underlying the centrality effect in memory, observing this effect simultaneously in memory and summarization datasets, as demonstrated in the present work, highlights the importance of considering the computational goal of these tasks. While the explicit goal of memory recall is to retrieve as much information as possible, in more realistic settings, recall is additionally constrained by limited time and cognitive resources. Under these conditions, it is rational for memory to prioritize central information, maximizing informativeness while minimizing effort. This explains why recall demonstrates patterns similar to those of summarization. This interpretation aligns with theories proposing that memory balances recall gains against costs (Anderson, 1990; Callaway et al., 2024; Ma et al., 2026; Van den Berg & Ma, 2018), and with modeling work showing that patterns of semantic memory search are well captured by algorithms like Google’s PageRank, which are designed to prioritize returning important information (Griffiths et al., 2007).

Another key advantage of our approach is that it unlocks existing summarization datasets as a rich and alternative resource for studying memory in naturalistic settings. Traditional memory experiments carried out in controlled laboratory settings, such as list learning paradigms (Kahana, 1996; Murdock, 1962), involve simplified, discrete stimuli. Although useful in uncovering fundamental memory mechanisms, they do not fully capture the range of dynamics and behaviors shown in real-world scenarios (Cohen & Conway, 2007; Huk et al., 2018). In response, recent studies have used more naturalistic settings to probe memory (Baldassano et al., 2018; Chen et al., 2017; Miller et al., 2013; Zacks et al., 2001) and incorporated social interactions important for real-world contexts (Angne et al., 2024;

Rajaram & Pereira-Pasarin, 2010; Weldon et al., 2000). However, these naturalistic approaches face a significant bottleneck: data scarcity. Particularly, naturalistic paradigms of memory yield a limited number of trials per experiment due to the nature of the naturalistic stimuli and procedures. To illustrate, consider the influential Sherlock dataset collected by Chen et al. (2017), which has supported analyses in over twenty studies of naturalistic memory (Hahamy et al., 2023; Heusser et al., 2021; Song et al., 2021). In this paradigm, collecting a single trial of data required a subject to watch a 50-minute movie, BBC’s Sherlock (initial portion of episode 1), and later to verbally recount what they could remember from the movie for up to 40 minutes – a commitment of nearly 90 minutes per trial. This high cost of data collection can limit the scale of naturalistic memory research. In contrast, the task of summarization offers far greater data availability, as large datasets of human-written summaries have been developed in the machine learning community used to evaluate automated models of summarization (for a review, see H. Zhang et al., 2025). These corpora are not only built on naturalistic stimuli but also reflect situations closer to how we encode and reconstruct information in our everyday lives. The present work serves as a proof-of-concept, demonstrating how these rich datasets can be leveraged to study the cognitive processes of both summarization and memory. Moving forward, they represent a valuable and largely untapped resource for cognitive science researchers. That said, as these datasets were not curated with psychology research in mind, there is less control over how the summaries were collected. Building on the results analyzed from these large datasets, it would be a fruitful avenue in the future to conduct laboratory-based studies that better control the texts, populations, and task instructions, such as whether the summarization task is given to subjects before or after encoding. Moreover, there are other important aspects of summarization that are not accounted for in the datasets analyzed here. Recent modeling work has shown that during narrative processing, such as listening to a story, the default mode network (DMN) in the brain constantly engages in online summarization of long-term context accumulated across minutes, hours, and even days (Tikochinski et al., 2025). Future empirical studies should examine how this more dynamic form of summary is maintained in working memory and how it changes over time during active narrative comprehension.

Lastly, our work helps bridge the study of summarization in psychology and machine learning. The present work has greatly benefited from the datasets developed in the machine

learning literature; in return, the insights uncovered from our study of summarization have implications for improving machine learning models of summarization. Since human summaries serve as gold standards for evaluating automated models of summarization (Over et al., 2007; T. Zhang et al., 2024), understanding the cognitive processes involved in human summarization could contribute to engineering better models of automated summarization. This understanding can not only improve evaluation metrics but is also useful in ensuring that generated outputs align with human expectations – an important consideration given that humans are the end users of these models (Mani, 2001). In line with this, several prior studies have leveraged cognitive theories to inform the design of summarization models (Cardenas et al., 2021; R. Zhang et al., 2016). Future summarization models could consider incorporating the memory processes identified in the current study, enabling them to generate summaries that better match human summarization patterns.

In conclusion, our findings contribute new insights to both memory and summarization research. We formally establish a relationship between the two tasks by showing where they align and where they diverge. This is facilitated through the use of tools and theories from both literatures, and by repurposing summarization datasets from machine learning for human behavioral analysis. By demonstrating the close link between summarization and memory recall, this work creates the foundation for developing integrative theories of how humans comprehend, retain, and reconstruct information.

Materials and methods

Dataset	Metadata*	Task demand	Example trial	
Narrative free recall	(843, 8–70, 1–55)	Subjects read source narrative as a scrolling text and immediately recall as much as possible from it by typing into a text box. There is no time or word limit on the recall phase.	Source: In the neighborhood I grew up in, Mrs. Baker was the resident cat lady. She had adopted at least...	Recall: There was a girl who lived next door to a woman who had 20 cats (or so) plus her son. Her son...
News articles	(302, 6–110, 1–4)	Subjects read a news article and are instructed to type a summary of the article in style of a newsletter of around 50 words.	Source: Children of the 1970s and 1980s will likely remember the chunky, plastic Fisher-Price...	Summary: Researchers are studying synaesthesia, a neurological condition where two...
HIPPOCORPUS	(2779, 13–34, 1–7)	Subjects are instructed to type an autobiographical experience from the recent past in 15 to 25 sentences, and to subsequently summarize in 1 to 3 sentences.	Source: It was a nice summer in Florida. The surf forecast was clean and 1-2 feet surf with offshore...	Summary: It was a nice day of surfing at the beach with my family. We all had a great time...
Reddit TIFU	(42139, 1–453, 1–23)	Subjects type an experience into a discussion forum along with a short summary of the experience.	Source: i ran track in high school and run with friends as a recreational activity. i'll get right...	Summary: decided to take a shortcut during my jog in a cemetery and fell in an open grave.

*(Dataset size, range of source material lengths, range of recall/summary lengths)

Table 1: An overview of datasets, including the Narrative free recall dataset (Georgiou et al., 2025) and three summarization datasets – News articles (T. Zhang et al., 2024), HIPPOCORPUS (Sap et al., 2022), and Reddit TIFU (Kim et al., 2019).

Datasets

To examine the relationship between memory recall and summarization, we analyzed four datasets from two types of tasks: the narrative free recall task (Georgiou et al., 2025), and the summarization task (Kim et al., 2019; Sap et al., 2022; T. Zhang et al., 2024). For each of these datasets, we filtered out trials that had empty source, recall, or summary fields. The metadata for the datasets is summarized in Table 1, along with task demands and example trials. We now describe these datasets.

Narrative free recall

The Narrative free recall dataset is taken from a large-scale online study (Georgiou et al., 2025), where 854 native English-speaking subjects were recruited through the Prolific platform. The study aimed to validate an automated pipeline for assessing memory performance, using large language model-generated narratives and automatic scoring of subjects' free recall. Narrative stimuli were generated using a large language model (GPT-4; Achiam et al., 2023) based on first-person oral narratives from Labov and Waletzky (1997) and Labov (2013), along with two original narratives from these sources. Subjects were first presented with a narrative in the form of rolling text at a rate of 12 characters per second.

Following this initial presentation, they were prompted to freely recall all possible details from the story by typing into a text box. The experiment concluded once subjects submitted their response with a button press (Georgiou et al., 2025).

News articles

The News articles dataset is taken from T. Zhang et al. (2024), where 6 freelance writers from Upwork were recruited after a qualification round where prospective writer summaries were assessed based on faithfulness, coherence, and relevance. The dataset was collected to address concerns about the poor quality of reference summaries commonly used to evaluate machine learning models and to compare the performance of human writers against large language models. 50 articles were selected from the evaluation sets of both the CNN/DailyMail (Nallapati et al., 2016) and XSum (Narayan et al., 2018) corpora to serve as stimuli. Each writer summarized a subset of these articles depending on their availability, with the most prolific writer summarizing 100 articles and the least summarizing 35. They were prompted to provide a summary of around fifty words in the style of a newsletter update (T. Zhang et al., 2024).

HIPPOCORPUS

The HIPPOCORPUS dataset was collected to investigate the use of natural language processing tools for studying human cognition (Sap et al., 2020, 2022). It includes data from 2,662 crowdworkers (age range 18 to 55 years old) recruited through Amazon Mechanical Turk. In the study, subjects recalled, imagined, or retold stories a few months after an initial recall phase. During the initial recall phase, subjects wrote a story about a memorable event they experienced in the past three to six months and then summarized it. We use data from this first recall phase for our analysis, where the recalled story acts as the source material. This phase includes 2,779 trials in total: 2,550 subjects contributed one story, 107 contributed two, and 5 contributed three.

Reddit TIFU

The Reddit TIFU dataset is taken from Kim et al. (2019) and was introduced to overcome patterns found in traditional summarization datasets, where summaries can often be constructed by copying sentences from predictable positions in the source. It contains data from the TIFU subreddit – a discussion forum containing posts of autobiographical stories – from January 2013 to March 2018. The title of each post is a short summary that

encapsulates the nature of the event described in the post (TIFU-short). Each post also ends with a longer summary that is descriptive of the post (TIFU-long). We use the TIFU-long version of the dataset, which contains 42,984 posts.

Mapping summary sentences to source sentences

The Narrative free recall and summarization datasets contain recalls and summaries that are not necessarily verbatim repetitions of sentences from the source material. Unlike standard free recall, where each recalled word can be directly matched to one on the presentation list, these datasets require identifying which source sentences are represented in a recall or summary. To address this, we first tokenized the source and summary sentences, then identified the best-matching source sentence for each summary (or recall) sentence using the maximum of a similarity metric, METEOR (Metric for Evaluation of Translation with Explicit Ordering) score (Banerjee & Lavie, 2005). METEOR computes similarity between a given sentence pair by accounting for word overlap, stemming, semantic alignment, and local word order. By going beyond exact word overlap to map semantic synonyms, METEOR ensures that similarity is captured even when source material is paraphrased during summarization. Originally developed for evaluating machine-generated translations, METEOR has also been widely adopted for assessing the performance of automated text summarization and has been shown to correlate well with human judgments of consistency, fluency, and relevance of generated summaries (Fabbri et al., 2021). Another consideration with the Narrative free recall and summarization datasets is the computation of serial position curves, given the variability in source lengths across trials. To enable aggregation, we normalized source sentence positions to a range of 0 to 1 and used bins – fixed-width intervals in this range – to average across trials. Specifically, we used 20 bins for the Narrative free recall, HIPPOCORPUS, and Reddit TIFU datasets and 40 bins for the News articles dataset, which contains a larger number of trials with longer source lengths.

Quantifying semantic centrality of source sentences

To quantify the semantic centrality of source sentences, we first transformed each source material into a weighted graph. In this representation, the source sentences, or nodes, form connections with each other via undirected edges. The strength of a connection between a pair of sentences, represented by the edge weight, is determined by the semantic similarity between them. After tokenizing sentences, this semantic similarity is computed using

METEOR score (Banerjee & Lavie, 2005). For a source with n sentences, we denote these semantic similarities using a $n \times n$ matrix L , where L_{ij} is the weighted edge from source sentence j to source sentence i (we have $L_{ij} = L_{ji}$ in the case of undirected edges). L is then transformed into a matrix M , where each source sentence distributes its weight proportionally over its outgoing edges, defined as $M_{ij} = L_{ij} / \sum_{k=1}^n L_{kj}$. The semantic centrality of a source sentence is defined as the PageRank of its corresponding node in the graph (Griffiths et al., 2007; Page et al., 1999). PageRank is based on the principle that important nodes receive links from other important nodes, implying a recursive definition of importance. Formally, this involves finding the principal eigenvector, p , of the matrix M that satisfies the equation $p = Mp$. By accounting for whether the neighbors of a node are themselves central, PageRank captures the global structure of the semantic graph. This global measure has been shown to outperform measures based on associate frequency – which assumes all word cues should be given equal weight – in predicting retrieval behavior in a related semantic fluency task (Griffiths et al., 2007).

Automated event segmentation of source material

Source material was segmented into events using the method from Kumar et al. (2023) that was found to best align with human judgments of event boundaries. We first obtained next-token probability distributions from the XL version of GPT-2 (Radford et al., 2019), which uses preceding context (words from previous time points) to predict the next token. At each position in the tokenized source, GPT-2 generated a probability distribution over its vocabulary. To estimate event boundaries, we calculated the Bayesian surprise at each token position as the KL-divergence between successive token distributions. The Bayesian surprise values for a source were then normalized with a running average of predictions using a drift parameter of 0.03. A peak in the Bayesian surprise values corresponds to a considerable shift in the model’s expectation of what token is next, indicating an event boundary in the source. Using these values and a percentile threshold, maxima above the chosen threshold were marked as event boundaries. We used four thresholds of 99%, 98%, 94%, and 90% to segment the source into events from large coarse-grained ones to progressively finer ones. The choice of the 90% threshold was consistent with a prior study on event segmentation (Michelmann et al., 2025), which showed alignment with aggregated human judgments of event boundaries. We included additional higher thresholds to capture more coarse-grained event segmentation.

Statistical analyses

In Analysis 1, permutation tests were performed for each dataset to determine whether the primacy and recency effects in overall recall and recall (or summary) initiation were significant. For overall recall, tests evaluated whether the recall probability of the target source sentence (first or last) was significantly greater than the recall probability of the middle three source sentences. For recall initiation, tests evaluated whether the proportion of recalls initiated by the target source sentence (first or last) was significantly greater than the proportion initiated by the middle three sentences. For each test, the position labels of binary recall outcomes for the target and middle positions were randomly shuffled 1,000 times, and p-values were computed as the proportion of permutations in which the difference in recalculated means was equal to or greater than the observed difference. For each dataset, we also performed permutation tests to determine whether the temporal contiguity and forward asymmetry effects in recall transitions were significant. For temporal contiguity, we tested whether the average conditional response probability (CRP) of nearby transitions (lags ± 1) was significantly greater than that of far away transitions (lags ± 2 through ± 7). For forward asymmetry, we evaluated whether the average CRP of forward transitions (lags $+1, +2, +3$) was significantly greater than that of backward transitions (lags $-1, -2, -3$). For each test, the sequence of recalled items within each trial was randomly shuffled 1,000 times, and p-values were computed as the proportion of permutations in which the difference in recalculated means was equal to or greater than the observed difference. In Analysis 2, a two-tailed Wilcoxon signed-rank test was conducted for each dataset. The test evaluated whether, across trials, sentences in the high semantic centrality group had significantly different recall probabilities compared to those in the low semantic centrality group. For each dataset, we also fit a logistic regression model using generalized estimating equations (GEE) to predict the recall status (recalled or not recalled) of a sentence. Semantic centrality group (high or low), sentence bin position, and their interaction were fixed effects, and observations were clustered by subjects. In Analysis 3, to quantify the strength of temporal contiguity in transition patterns for different event segmentation thresholds, we fit linear regression models predicting conditional response probability as a function of positive transition lags ($+1$ to $+7$). Separate models were estimated for each dataset and each segmentation threshold (99%, 98%, 94%, and 90%). For the finer thresholds of 94% and 90%, six permutation tests were performed to

determine whether the slope differences between the Narrative free recall dataset and each summarization dataset were significant. For each test, trial-level dataset labels were randomly shuffled 1,000 times, and p-values were computed as the proportion of permutations in which the difference in recalculated slopes was equal to or greater than the observed difference.

Transparency and Openness

Publicly available datasets were used in this study. The analysis code and datasets used can be accessed here: <https://osf.io/bx2n6> (Angne, 2026). The study was not preregistered.

Acknowledgments

This work was supported by a start-up fund awarded to Q.Z. by Rutgers University–New Brunswick. This work was also supported by the National Science Foundation’s National Research Traineeship (2021628) awarded to H. A.. The authors declare no competing interests.

Constraints on Generality

The reported findings use datasets that span multiple task demands and narrative types, including summaries of news articles, autobiographical accounts, informal online posts, and recalls of first-person narratives. In regard to demographics, the Narrative free recall dataset (Georgiou et al., 2025) recruited native English speakers (no additional demographic information was collected). The Martinez (2024) dataset recruited native English speakers aged 18 to 35 through the Prolific platform and the University of Maryland psychology participant pool, with 70% identifying as White and 49.8% as female, with other groups represented in smaller proportions. The Racciah et al. (2024) dataset recruited native English speakers aged 17 to 75 through the SONA Systems platform at New York University and the Prolific platform, which included 229 subjects, out of which 145 identified as female. The News articles dataset (T. Zhang et al., 2024) consists of summaries written by writers from the freelance work platform Upwork, with prior experience writing professional content. The HIPPOCORPUS dataset (Sap et al., 2020, 2022) recruited story authors aged 18 to 55, 73.7% of whom identified as White and 52% as female, with other groups represented in smaller proportions. Finally, the Reddit TIFU dataset (Kim et al., 2019) consists of user-generated content from a large online community, but no demographic information is available about its contributors. The observed recall and summarization patterns generalized across these datasets, suggesting that the patterns represent salient features of how we reconstruct information across different task demands, narrative types, and populations. Nevertheless, it remains unclear how these results extend to other retrieval instructions or reader goals, which can influence what information is encoded and subsequently recalled or summarized.

References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.
- Angne, H. (2026, June). Bridging real-world summarization and laboratory-based memory recall. <https://doi.org/10.17605/OSF.IO/BX2N6>
- Angne, H., Cornell, C. A., & Zhang, Q. (2024). A context-based model of collaborative inhibition during memory search. *Scientific Reports*, 14(1), 27645.
- Antony, J., Lozano, A., Dhoat, P., Chen, J., & Bennion, K. (2024). Causal and chronological relationships predict memory organization for nonlinear narratives. *Journal of Cognitive Neuroscience*, 36(11), 2368–2385.
- Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, 38(45), 9689–9699.
- Banerjee, S., & Lavie, A. (2005). Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.
- Brown, A. L., Day, J. D., & Jones, R. S. (1983). The development of plans for summarizing texts. *Child development*, 968–979.
- Callaway, F., Griffiths, T. L., Norman, K. A., & Zhang, Q. (2024). Optimal metacognitive control of memory recall. *Psychological Review*, 131(3), 781–811.
- Cardenas, R., Galle, M., & Cohen, S. B. (2021). Unsupervised extractive summarization by human memory simulation. *arXiv preprint arXiv:2104.08392*.
- Carvalho, W., & Lampinen, A. (2025). Naturalistic computational cognitive science: Towards generalizable models and theories that capture the full range of natural behavior. *arXiv preprint arXiv:2502.20349*.
- Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, 20(1), 115–125.
- Cohen, G., & Conway, M. A. (2007). *Memory in the real world*. Psychology Press.

- Cornell, C. A., Norman, K. A., Griffiths, T. L., & Zhang, Q. (2024). Improving memory search through model-based cue selection. *Psychological Science*, 35(1), 55–71.
- Denhiere, G., Dessus, P., Lemaire, B., & Mandin, S. (2005). Computational cognitive models of summarization assessment skills. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 27(27), 1266–1271.
- Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22, 457–479.
- Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9, 391–409.
- Georgiou, A., Can, T., Katkov, M., & Tsodyks, M. (2025). Large-scale study of human memory for meaningful narratives. *Learning & Memory*, 32(2), a054043.
- Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with pagerank. *Psychological science*, 18(12), 1069–1076.
- Hahamy, A., Dubossarsky, H., & Behrens, T. E. (2023). The human brain reactivates context-specific past information at event boundaries of naturalistic experiences. *Nature neuroscience*, 26(6), 1080–1089.
- Heusser, A. C., Fitzpatrick, P. C., & Manning, J. R. (2021). Geometric models reveal behavioural and neural signatures of transforming experiences into memories. *Nature Human Behaviour*, 5(7), 905–919.
- Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(4), 923–941.
- Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3), 269–299.
- Huk, A., Bonnen, K., & He, B. J. (2018). Beyond trial-based paradigms: Continuous behavior, ongoing neural activity, and natural stimuli. *Journal of Neuroscience*, 38(35), 7551–7558.
- Jeunehomme, O., & D’Argembeau, A. (2020). Event segmentation and the temporal compression of experience in episodic memory. *Psychological research*, 84(2), 481–490.

- Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24(1), 103–109.
- Kahana, M. J. (2012). *Foundations of human memory*. OUP USA.
- Kim, B., Kim, H., & Kim, G. (2019). Abstractive summarization of reddit posts with multi-level memory networks. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2519–2531.
- Kintsch, W. (2008). On the notions of theme and topic in psychological process models of text comprehension. In *Thematics: Interdisciplinary studies* (pp. 157–170). John Benjamins Publishing Company.
- Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review*, 85(5), 363–394.
- Kumar, M., Goldstein, A., Michelmann, S., Zacks, J. M., Hasson, U., & Norman, K. A. (2023). Bayesian surprise predicts human event segmentation in story listening. *Cognitive science*, 47(10), e13343.
- Labov, W. (2013). *The language of life and death: The transformation of experience in oral narrative*. Cambridge University Press.
- Labov, W., & Waletzky, J. (1997). Narrative analysis: Oral versions of personal experience1. *Journal of Narrative and Life History*, 7(1-4), 3–38.
- Lee, H., & Chen, J. (2022). Predicting memory from the network structure of naturalistic events. *Nature Communications*, 13(1), 4235.
- Lehnert, W. G. (2014). Plot units: A narrative summarization strategy. In *Strategies for natural language processing* (pp. 375–412). Psychology Press.
- Ma, S., Popov, V., & Zhang, Q. (2026). Cognitive constraints and reward environments jointly shape memory formation. *Psychological Review*.
- Mandler, G. (2011). *A history of modern experimental psychology: From james and wundt to cognitive science*. MIT press.
- Mani, I. (2001). *Automatic summarization*. John Benjamins Publishing Company.
- Martinez, D. (2024). Scoring story recall for individual differences research: Central details, peripheral details, and automated scoring. *Behavior Research Methods*, 56(8), 8362–8378.

- McNamara, D. S., Watanabe, M., Huynh, L., McCarthy, K. S., Allen, L. K., & Magliano, J. P. (2024). Summarizing versus rereading multiple documents. *Contemporary Educational Psychology*, 76, 102238.
- Michelmann, S., Kumar, M., Norman, K. A., & Toneva, M. (2025). Large language models can segment narrative events similarly to humans. *Behavior Research Methods*, 57(1), 1–13.
- Miller, J. F., Neufang, M., Solway, A., Brandt, A., Trippel, M., Mader, I., Hefft, S., Merkow, M., Polyn, S. M., Jacobs, J., et al. (2013). Neural activity in human hippocampal formation reveals the spatial context of retrieved memories. *Science*, 342(6162), 1111–1114.
- Murdock, B. B. (1960). The distinctiveness of stimuli. *Psychological Review*, 67(1), 16–31.
- Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5), 482–488.
- Musz, E., & Chen, J. (2022). Neural signatures associated with temporal compression in the verbal retelling of past events. *Communications biology*, 5(1), 489.
- Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, 280–290.
- Narayan, S., Cohen, S. B., & Lapata, M. (2018). Don’t give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1797–1807.
- Omanson, R. C. (1982). The relation between centrality and story category variation. *Journal of Verbal Learning and Verbal Behavior*, 21(3), 326–337.
- Ophuis-Cox, F., Rozendal, L., Catrysse, L., Joosten-ten Brinke, D., & Camp, G. (2024). The effects of summarization and factual retrieval practice on text comprehension and text retention in elementary education. *Journal of Experimental Psychology: Applied*, 30(2), 258.
- Over, P., Dang, H., & Harman, D. (2007). Duc in context. *Information Processing & Management*, 43(6), 1506–1520.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web*. (tech. rep.). Stanford infolab.

- 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.
- Racah, O., Chen, P., Gureckis, T. M., Poeppel, D., & Vo, V. A. (2024). The “naturalistic free recall” dataset: Four stories, hundreds of participants, and high-fidelity transcriptions. *Scientific Data*, 11(1), 1317.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8).
- Rajaram, S., & Pereira-Pasarin, L. P. (2010). Collaborative memory: Cognitive research and theory. *Perspectives on Psychological Science*, 5(6), 649–663.
- Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology*, 92(3), 365–372.
- Rumelhart, D. E. (1977). Understanding and summarizing brief stories. *Basic processes in reading: Perception and comprehension*, 265, 303–343.
- Russek, E. M., Acosta-Kane, D., van Opheusden, B., Mattar, M. G., & Griffiths, T. L. (2025). Time spent thinking in online chess reflects the value of computation. *Cognitive science*, 49(10), e70119.
- Sap, M., Horvitz, E., Choi, Y., Smith, N. A., & Pennebaker, J. (2020). Recollection versus imagination: Exploring human memory and cognition via neural language models. *Proceedings of the 58th annual meeting of the association for computational linguistics*, 1970–1978.
- Sap, M., Jafarpour, A., Choi, Y., Smith, N. A., Pennebaker, J. W., & Horvitz, E. (2022). Quantifying the narrative flow of imagined versus autobiographical stories. *Proceedings of the National Academy of Sciences*, 119(45), e2211715119.
- Song, H., Finn, E. S., & Rosenberg, M. D. (2021). Neural signatures of attentional engagement during narratives and its consequences for event memory. *Proceedings of the National Academy of Sciences*, 118(33), e2021905118.
- Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, 25(2), 207–222.
- Steyvers, M., & Schafer, R. J. (2020). Inferring latent learning factors in large-scale cognitive training data. *Nature Human Behaviour*, 4(11), 1145–1155.

- Taieb-Maimon, M., Romanovski-Chernik, A., Last, M., Litvak, M., & Elhadad, M. (2024). Mining eye-tracking data for text summarization. *International Journal of Human-Computer Interaction*, 40(17), 4887–4905.
- Tan, L., Ward, G., Paulauskaite, L., & Markou, M. (2016). Beginning at the beginning: Recall order and the number of words to be recalled. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(8), 1282–1292.
- Tikochinski, R., Goldstein, A., Meiri, Y., Hasson, U., & Reichart, R. (2025). Incremental accumulation of linguistic context in artificial and biological neural networks. *Nature communications*, 16(1), 803.
- Van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *ELife*, 7, e34963.
- Weldon, M. S., Blair, C., & Huesch, P. D. (2000). Group remembering: Does social loafing underlie collaborative inhibition? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(6), 1568–1577.
- Winograd, P. N. (1984). Strategic difficulties in summarizing texts. *Reading research quarterly*, 404–425.
- Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of experimental psychology: General*, 130(1), 29–58.
- Zhang, H., Yu, P. S., & Zhang, J. (2025). A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11), 1–41.
- Zhang, Q. (2026). Formal modeling as theoretical glue between laboratory and naturalistic studies of memory. *Perspectives on Psychological Science*, 0(0), 17456916261419428.
- Zhang, Q., Griffiths, T. L., & Norman, K. A. (2023). Optimal policies for free recall. *Psychological Review*, 130(4), 1104–1124.
- Zhang, R., Li, W., Liu, N., & Gao, D. (2016). Coherent narrative summarization with a cognitive model. *Computer Speech & Language*, 35, 134–160.
- Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12, 39–57.

References

- Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., ... others (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.
- Georgiou, A., Can, T., Katkov, M., & Tsodyks, M. (2025). Large-scale study of human memory for meaningful narratives. *Learning & Memory*, 32(2), a054043.
- Lee, H., & Chen, J. (2022). Predicting memory from the network structure of naturalistic events. *Nature communications*, 13(1), 4235.
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).
- Martinez, D. (2024). Scoring story recall for individual differences research: Central details, peripheral details, and automated scoring. *Behavior Research Methods*, 56(8), 8362–8378.
- Racah, O., Chen, P., Gureckis, T. M., Poeppel, D., & Vo, V. A. (2024). The “naturalistic free recall” dataset: four stories, hundreds of participants, and high-fidelity transcriptions. *Scientific Data*, 11(1), 1317.

Supplementary Materials

S1. Alternative methods for mapping summary sentences to source sentences

In this section, we apply two alternative similarity metrics for mapping summary or recall sentences to a sentence in the source. This allows us to test whether our conclusion regarding the temporal dynamics in summarization aligning with memory recall holds under different mapping methods. The first alternative metric is the ROUGE-1 (Recall-Oriented Understudy for Gisting Evaluation) score. ROUGE-1 is used to evaluate the performance of automated text summarization. It measures similarity based on the number of overlapping unigrams – single words or tokens – between the generated summary and a set of gold reference summaries (Lin, 2004). The second metric is the cosine similarity between sentence embeddings, where the sentences are embedded into a 512-dimensional semantic space using Google’s Universal Sentence Encoder (USE; Cer et al., 2018).

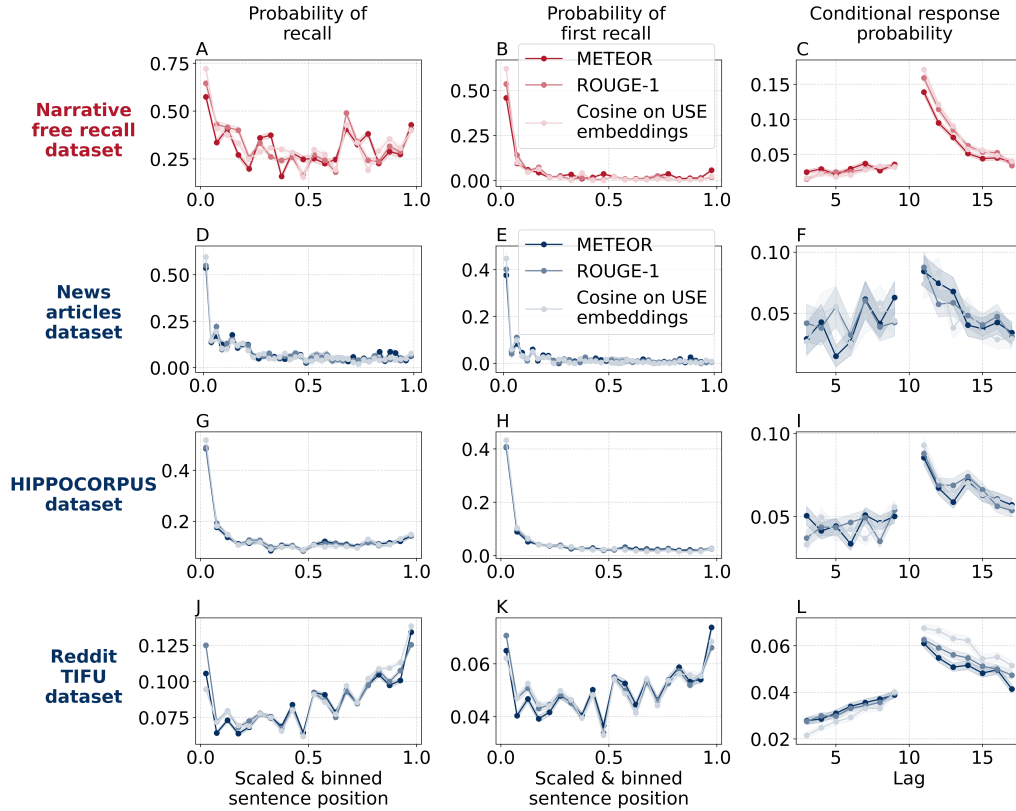


Figure S1. Temporal dynamics in summarization align with patterns observed in memory recall, irrespective of the mapping method used. Each sentence in the summary was mapped to a sentence in the source based on three similarity metrics – METEOR score, ROUGE-1 score, and cosine between USE embeddings of sentences. The behavioral patterns of the serial position curve, probability of first recall, and the conditional response probability showed consistent trends when plotted using each of the three metrics for the (A–C) Narrative free recall, (D–F) News articles, (G–I) HIPPOCORPUS, and (J–L) Reddit TIFU datasets. The shaded error represents the standard error of the mean.

This metric has been previously used in memory recall literature to capture the semantic content of sentences (Lee & Chen, 2022). For both metrics, we calculated the similarity between a summary sentence and each source sentence, identifying the best-matching source sentence using the maximum of the ROUGE-1 score or USE cosine similarity value. We analyzed the behavioral patterns of the serial position curve, the probability of first recall, and the conditional response probability, comparing mappings obtained from the METEOR score, ROUGE-1 score, and USE cosine similarity. In Figure S1, different mapping methods showed similar behavioral patterns, supporting that our conclusions are not sensitive to the specific mapping method used.

S2. Additional naturalistic recall datasets

We analyzed two additional naturalistic recall datasets from Martinez (2024) and Raccach, Chen, Gureckis, Poeppel, and Vo (2024), to examine whether recall and summarization behaviors observed in the primary Narrative free recall dataset (Georgiou, Can, Katkov, & Tsodyks, 2025) generalize to other naturalistic contexts and populations. Both additional datasets use continuous, semantically related source material. We followed the same analyses we used for the primary dataset, specifically examining temporal dynamics, centrality effects, and transition patterns at multiple hierarchies of the source.

Martinez (2024) dataset: This dataset is taken from a study in Martinez (2024), where 235 subjects (age range 18 to 35 years old) were recruited online through the Prolific platform or the psychology research participant pool at the University of Maryland. We use data from the first two story recall tasks in Martinez (2024). A third task was excluded as it involved encoding multiple stories before recall. In the first task, subjects were presented with a story between 62 and 77 words on the screen for 45 seconds. Following this, they were presented with a filler task and then had up to 2 minutes to type their recall in a text box. Each subject completed this procedure for three stories. The second task followed a similar paradigm, except the source material was presented aurally for about 21 to 25 seconds.

Raccach et al. (2024) dataset: The Raccach et al. (2024) dataset contains transcribed verbal recollections of four spoken narratives collected from 229 subjects (age range from 17 to 75 years old). The data was collected on two online platforms: New York University’s SONA Systems platform and Prolific. The study paradigm involved each subject listening to two stories of length 8 to 13 minutes. Immediately after each story, subjects were asked to verbally recall the source content in as much detail as possible for a minimum of 4 minutes. While subjects were encouraged to recall the story in the correct order, they were asked to return to any earlier points they might have missed. Verbal recalls were transcribed using Google Cloud Speech-to-Text and manually verified against the original audio files by a commercial transcript correction service (Raccach et al., 2024).

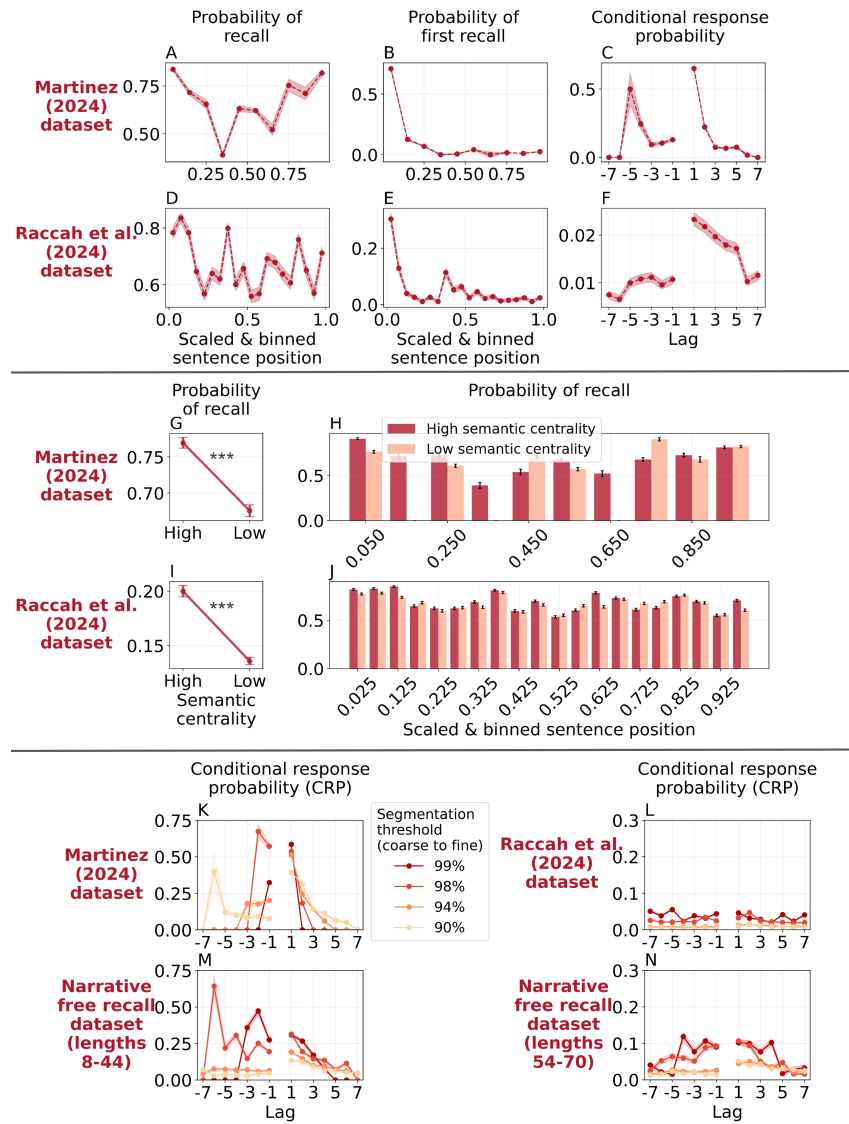


Figure S2. Additional datasets of naturalistic recall show memory recall and summarization effects. Temporal dynamics in (A–C) the Martinez (2024) and (D–F) the Raccach et al. (2024) datasets align with standard free recall effects of primacy, recency, and temporal contiguity. Semantically central source sentences were more likely to be recalled, and this advantage was centered around early source sentences for both (G–H) the Martinez (2024) and (I–J) the Raccach et al. (2024) datasets. Conditional response probability curves at four segmentation thresholds were plotted for (K) the Martinez (2024) and (L) the Raccach et al. (2024) datasets. The Martinez (2024) dataset showed steep negative slopes at all segmentation thresholds, similar to the trend observed in (M) the short source length trials of the Narrative free recall dataset. The Raccach et al. (2024) dataset showed steep negative slopes only at the coarse-grained segmentation thresholds, with flatter slopes at finer thresholds of 94% and 90%, similar to the trend observed in (N) the long source length trials of the Narrative free recall dataset. The shaded error represents the standard error of the mean.

Analysis 1. To test whether primacy and recency were significant in these additional datasets, we compared the likelihood of recalling or initiating a recall from the first or last source sentence against the middle three. Subjects in both the naturalistic recall datasets showed a primacy effect in overall recall ($p < .001$ for all comparisons, determined by non-parametric permutation tests in which the position labels of binary recall outcomes were randomly shuffled 1,000 times separately for each dataset; Figures S2AD), as well as a recency effect ($p < .001$ for all comparisons, determined by the same permutation tests; Figures S2AD). Recall initiation showed a primacy effect for both datasets ($p < .001$ for all comparisons, determined by the same permutation tests; Figures S2BE) and a recency effect for the Martinez (2024) dataset ($p = .011$, determined by the same permutation test; Figure S2B). Subjects in both datasets showed a temporal continuity effect, with more nearby transitions (at lags ± 1) than farther away ones (at lags ± 2 through ± 7) ($p < .001$ for all comparisons; determined by non-parametric permutation tests in which recall sequences were randomly shuffled 1,000 times; Figures S2CF). They also showed a forward asymmetry effect, with a greater tendency for forward transitions (at lags $+1, +2, +3$) than backward ones (at lags $-1, -2, -3$) ($p < .001$ for all comparisons; determined by the same permutation tests). These temporal dynamics align with standard free recall patterns and the patterns observed in the Narrative free recall dataset.

Analysis 2. In both additional datasets, semantically central source sentences were more likely to be recalled, replicating the centrality effect observed in the primary Narrative free recall and summarization datasets. Specifically, the probability of recall was greater for the high semantic centrality group than the low semantic centrality group (Figures S2GI) and this difference was significant for the Martinez (2024) dataset (Wilcoxon signed-rank test:

$W = 98592.5, n = 1148, p < .001$, two-tailed), and the Racciah et al. (2024) dataset (Wilcoxon signed-rank test: $W = 13266.0, n = 458, p < .001$, two-tailed). Both datasets showed that this effect was significant across sentence position bins (Martinez (2024) dataset

OR = 1.53, 95%CI [1.29, 1.82], $p < .001$; Racciah et al. (2024) dataset

OR = 1.19, 95%CI [1.12, 1.27], $p < .001$; Figures S2HJ). This centrality advantage diminished for sentences appearing later in the source, as indicated by a significant negative interaction between semantic centrality group and bin position (Martinez (2024) dataset

OR = 0.92, 95%CI [0.89, 0.95], $p < .001$; Racciah et al. (2024) dataset

OR = 0.99, 95%CI [0.985, 0.997], $p < .001$; Figures S2HJ). Both these datasets demonstrated centrality effects characteristic of summarization, strengthening the conclusion that recall behavior takes the form of summarization under naturalistic conditions.

Analysis 3. We next examined transition patterns at multiple hierarchies of the source. Both datasets showed temporal continuity at the coarse-grained segmentation thresholds of 99% and 98% (Figures S2KL). At fine-grained thresholds of 94% and 90%, the Martinez (2024) dataset continued to

exhibit strong temporal contiguity, aligning with the behavior of the primary Narrative free recall dataset. In contrast, the Racciah et al. (2024) dataset showed flatter slopes at these finer thresholds, aligning with behavior in the summarization datasets. One major difference between the Martinez (2024) dataset and the Racciah et al. (2024) dataset is the length of the source material, with the former ranging from 5 to 7 sentences and the latter from 72 to 190 sentences. To test whether source length modulates fine-grained temporal contiguity, we split the primary Narrative free recall dataset into trials with short and long source lengths and repeated the same analysis. At the finer thresholds of 94% and 90%, the Martinez (2024) dataset (source lengths of 5-7 sentences) and the short source trials from the primary Narrative free recall dataset showed steep negative slopes (Figures S2KM), with slope differences larger than those in the three summarization datasets ($p < .001$ for all comparisons; determined by non-parametric permutation tests with 1,000 shuffles separately for each naturalistic recall and summarization dataset pairing, and for each of the two thresholds). Conversely, the Racciah et al. (2024) dataset (source lengths of 72-190 sentences) and long source trials of the primary Narrative free recall dataset showed flattened slopes at these finer thresholds (Figures S2LN), similarly to those in the three summarization datasets ($p > 0.05$ for comparisons; determined by the same permutation tests). These results show that while memory recall and summarization share common characteristics, the hierarchical level at which information is reconstructed in each task is driven by both the task objective and the length of the source.