Summarization Reflects Characteristics of Memory Recall

Hemali Angne [a], **Snigdha S. Mishra** [a], **Guo Lin** [a], **Qiong Zhang** [a,b,c]
{h.angne, sm2600, gl550, qiong.z}@rutgers.edu
[a]Department of Computer Science, [b]Department of Psychology, [c]Rutgers Center for Cognitive Science
Rutgers University–New Brunswick, Piscataway, NJ, USA 08854

Abstract

Memory has traditionally been studied in well-controlled laboratory environments, which, while effective, do not fully capture the range of dynamics and behaviors shown in real-world contexts. To address this gap, we propose using summarization as a novel task to study memory recall in naturalistic settings. We argue that a key component of summarization is the ability to represent and retain information from the original material. Inspired by approaches in the free recall literature to analyze temporal dynamics of memory recall, such as how recall begins and transitions to subsequent items, we analyzed the temporal dynamics of summary patterns. Using three publicly available summarization datasets and a naturalistic narrative recall dataset, we found alignments between the summarization patterns and established free recall patterns, including primacy, recency, temporal contiguity, and the effect of list length. These results support that summarization involves processes of memory recall and open up opportunities to use summarization as a naturalistic task to study memory recall in the future.

Keywords: naturalistic recall; summarization; free recall task; memory bias

Introduction

Memory experiments carried out in controlled laboratory settings, such as list learning paradigms (Murdock, 1962; Kahana, 1996), involve simplified, discrete stimuli and do not fully capture the range of dynamics and behaviors shown in real-world scenarios (Cohen & Conway, 2007; Huk, Bonnen, & He, 2018). To address this limitation, recent studies have used more naturalistic settings to probe memory (Chen et al., 2017; Coutanche, Koch, & Paulus, 2020; Baldassano, Hasson, & Norman, 2018; Zacks, Tversky, & Iyer, 2001; Miller et al., 2013; Cornell, Jin, & Zhang, 2023) and incorporated social interactions important for real-world contexts (Weldon, Blair, & Huebsch, 2000; Rajaram & Pereira-Pasarin, 2010; Angne, Cornell, & Zhang, 2024). In the current work, we propose summarization as a novel task for studying memory recall in naturalistic settings. A summarization task not only involves naturalistic stimuli such as texts and narratives but also reflects situations closer to how we recall information in our everyday lives. Imagine a scenario where we ask a friend what they remember from a book they read. Constrained by limited time and effort in real life, they are more likely to provide us with a summary of the book rather than listing every single detail they can remember.

While recall in real-world scenarios often resembles summarization, we will formally analyze the relationship between summarization and recall in this study. Since the objective of summarization is to retain key information from the source material before such information is condensed or summarized, we hypothesize that a summarization task should first be considered a task of memory recall. In other words, one's ability to provide good summaries is constrained by their ability to recall all the information from the original source material. If this is the case, we should expect to see alignment between patterns in summarization data and known patterns previously identified in memory recall literature.

Specifically, when participants are asked to recall all the information they can remember from a list of words (in a free recall task), there are typical patterns associated with the temporal dynamics of their memory recalls, such as which word positions are more likely to be recalled and where one's recall is likely to start from (relative to the source material), as well as where one's next recall is likely to transit to (Murdock, 1962; Kahana, 1996). Inspired by this literature in examining the temporal dynamics of memory recalls, we will carry out the same analyses to understand the temporal dynamics of summarization. We analyze three summarization datasets – the News articles dataset (Zhang et al., 2024), the HIPPOCORPUS dataset (Sap et al., 2022), and the Reddit TIFU dataset (Kim, Kim, & Kim, 2019). 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, Can, Katkov, & Tsodyks, 2023), which is a naturalistic version of the standard free recall task, involving continuous, semantically related source material.

To foreshadow our results, our analysis of temporal dynamics reveals that there is an alignment in summarization patterns and standard memory recall patterns, suggesting that the way we summarize is constrained by our memory. This validates the use of summarization datasets to study naturalistic recall. For the rest of the paper, we first describe the datasets we use. We then review relevant analyses and findings from free recall literature and the techniques we use to similarly analyze narrative free recall and summarization datasets. Finally, we present our findings on patterns seen in summarization and discuss the implications of these results.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Dataset	Metadata*	Task demand	Example trial	
Narrative free recall	(843, 8–70, 1–55)	Recall as closely as possible to the source	**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)	Summarize article in style of a newsletter	**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)	Summarize an event from the past six months	**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)	Summarize an event descriptively	**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., 2023) and three summarization datasets – News articles (Zhang et al., 2024), HIPPOCORPUS (Sap et al., 2022), and Reddit TIFU (Kim et al., 2019).

Method

Narrative Free Recall and Summarization Datasets

To examine the role of recall in summarization, we analyze data from two types of tasks: the narrative free recall task – a naturalistic version of the standard free recall task, and the summarization task. We now describe these datasets below.

Narrative free recall The Narrative free recall dataset is from an online study in Georgiou et al. (2023), where subjects first studied source material presented in the form of rolling text with a speed of 25 characters per second. Following this, the recall phase consisted of a textbox and a prompt to recall the material as closely as possible to the original one (Georgiou et al., 2023).

News articles The News articles dataset is from Zhang et al. (2024), where six freelance writers were recruited after their summaries were assessed based on faithfulness, coherence, and relevance. These selected writers provided summaries for a limited set of samples from the CNN/DailyMail (Nallapati, Zhou, dos Santos, Gulçehre, & Xiang, 2016) and XSum (Narayan, Cohen, & Lapata, 2018) corpora in the style of a newsletter update (Zhang et al., 2024).

HIPPOCORPUS The HIPPOCORPUS dataset is from Sap et al. (2022) and contains data collected through an online study, where subjects recalled, imagined, or retold stories a few months after first recalling them. During the first recall phase, subjects wrote a story about a memorable event they experienced in the past 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.

Reddit TIFU The Reddit TIFU dataset is from (Kim et al., 2019). 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.

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. For the summarization datasets, information about whether subjects had access to the source material during summarization is not available. However, given one's limited working memory (Miller, 1956; Cowan, 2010), it is unlikely that subjects could hold the entirety of the source in mind at once. Therefore, even if they had opportunities to view the source material during summarization, memory retrieval processes would still be involved.

Our proposed analyses of temporal dynamics

We propose that the task of summarization can be analyzed as a task of memory recall, as one needs to retain key information from the source material and later cover it in a summary. We will first describe the temporal dynamics when recalling lists of words before we introduce how we are going to derive them similarly for summarization and narrative free recall.

Temporal dynamics analyzed in a free recall task The memory task in the literature that is most aligned with the goal of summarization is the 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). The free recall task involves subjects first studying a list of items, such as words, and then being asked to recall as many items as they can in any order from the list. One can consider this task as a simplified version of a real-life recall scenario, where one tries to remember all the information from a holiday event or a movie. We will refer to this task as the standard free recall task from now on to distinguish it from versions of the free

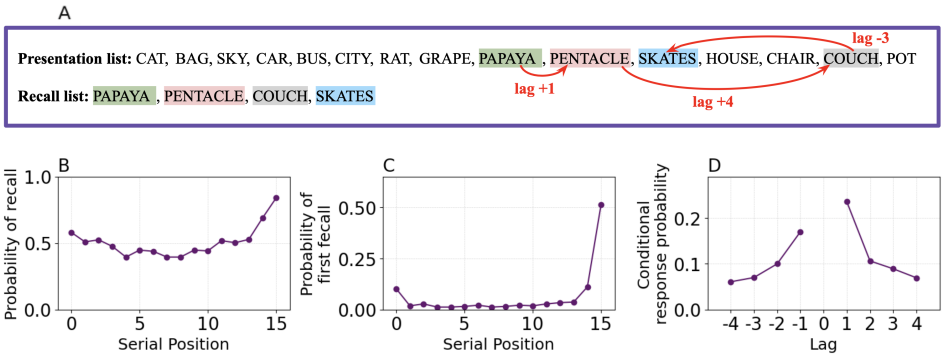

Figure 1: The free recall task. (A) In this illustrative example of the standard free recall task, the goal is to recall as many items as possible from the presentation list in any order. Items recalled at consecutive positions correspond to items studied either nearby (e.g., lag +1) or far away (e.g., lag -3, lag +4). Averaging across multiple trials gives patterns of temporal dynamics, including (B) the serial position curve, (C) the probability of first recall, and (D) the conditional response probability, plotted based on the dataset in Cornell et al. (2024).

recall task where the study material is continuous narratives (Narrative free recall; Georgiou et al., 2023).

An illustrative trial of this task is given in Figure 1A, where fifteen items were presented, and four items from the list were recalled. There is rich information in the order of recalled items. First, one can analyze which item positions in the presentation list 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). The standard free recall trial illustrated in Figure 1A provides an example where the first recall 'PAPAYA' is at position 9 in the presentation list. Aggregated across multiple trials, Figures 1B and 1C plot the serial position and probability of first recall curves from a representative 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 of the list or the end of the list (primacy or recency effects; Murdock, 1962). The **probability of first recall curve** demonstrates a similar tendency to either initiate recall from the beginning of the presentation list or the end. Both primacy and recency effects are robust across multiple studies, though their relative ratio can vary given task conditions (Howard & Kahana, 1999; Tan, Ward, Paulauskaite, & Markou, 2016) and how well subjects perform the task (Zhang, Griffiths, & Norman, 2023).

In addition to recall probability and initiation, one can also examine recall transitions by analyzing how likely it is to recall items studied consecutively in the presentation list (**conditional response probability**, computed by dividing the number of times a transition for a lag is actually made by the number of times it could have been made; Kahana, 1996). The standard free recall trial illustrated in Figure 1A provides an example where the lag is +1 going from the first recall 'PAPAYA' (at serial position 9) to the second recall 'PENTACLE' (at serial position 10), and the lag is +4 going from the second recall 'PENTACLE' (at serial position 10) to the third recall 'COUCH' (at serial position 14). Aggregated across multiple trials from the same study in Cornell et al. (2024), the conditional response probability in Figure 1D demonstrates a greater tendency of subjects to transit to a next recall that is at consecutive positions in the presentation list with smaller lags than farther away positions with larger lags (temporal contiguity effects; Kahana, 1996). In addition, subjects make both backward transitions (negative lags) and forward transitions (positive lags), though with a greater tendency for forward transitions (forward asymmetry; Kahana, 1996).

We additionally examine the effect of list length observed in free recall literature – the tendency of subjects to initiate recall at the first serial position decreases with increasing presentation list length (Ward, Tan, & Grenfell-Essam, 2010; Tan et al., 2016). These well-documented patterns in the task of standard free recall provide a framework for analyzing similar dynamics in narrative free recall and summarization data.

Temporal dynamics analyzed in narrative free recall and summarization tasks We will now describe how we analyze the narrative free recall and summarization datasets to identify the same patterns analyzed in the standard free recall task. In the narrative free recall and summarization tasks, sentences in the source material are analogous to items in a presentation list, while sentences in the recall or summary are analogous to recalled items. Similarly to how in the standard free recall task a recalled item is mapped to one of the items in the presentation list, we identify for each sentence in the summary or narrative recall which sentence in the source material it best maps to (or summarizes from) based on the similarity between them. We first tokenize each sentence in the summary (or recall) and the source material. Then, for each summary sentence, we identify a single sentence in the source as its best match using the maximum of a similarity metric, METEOR score, which accounts for word overlap, stemming, and semantic alignment (Banerjee & Lavie, 2005). For example, the first sentence in the summary, highlighted in green

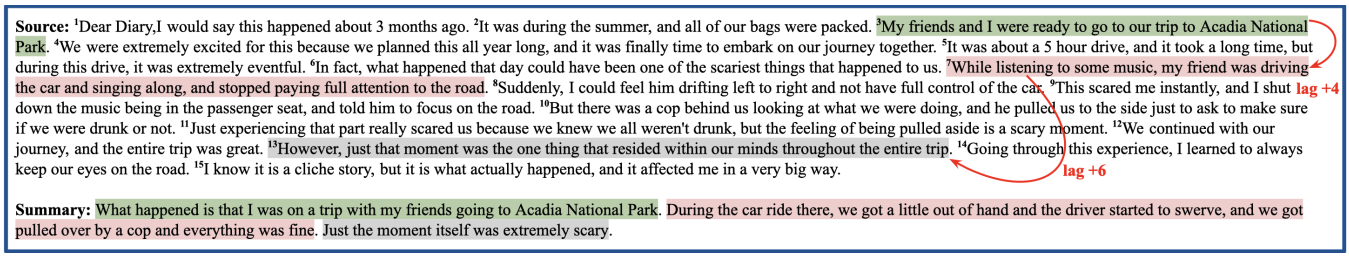

Figure 2: A trial of the summarization task from the HIPPOCORPUS dataset. In the summarization task, the goal is to retain key information from the source material. We identify, for each sentence in the summary, which sentence in the source material it best maps to (or summarizes from) based on a similarity metric, the METEOR score. Consecutive sentences in a summary correspond to either nearby (e.g., lag +4) or far away (e.g., lag +6) sentences in the source material.

in Figure 2, 'What happened is that I was on a trip with my friends going to Acadia National Park' is best aligned with the third sentence in the source material 'My friends and I were ready to go to our trip to Acadia National Park'.

Formulated this way, we can obtain the **serial position curve** by estimating which sentences in the source material are more likely to be included in the summary, analogous to Figure 1B. We can obtain the **probability of first recall** by estimating where the summary is likely to start from relative to the source material, analogous to Figure 1C. We normalize sentence positions in the source material to a range of 0 to 1 and use bins – fixed-width intervals in this range – to average across trials, enabling us to plot source material of varying lengths together. In addition, we can similarly derive the **conditional response probability** to examine the transition patterns (in lags) as a summary unfolds. In the example in Figure 2, the first and second summary sentences are mapped to position 3 and position 7 in the source material. So, the lag between the first and second summary sentence mappings is +4. Lastly, we introduce an **information score** measure to understand where important information is located in the source material. The information score is a set of METEOR score values calculated for each sentence in the source material between that sentence and the entire source. It estimates the importance of each sentence in the context of the source material. When items in a list are randomly assigned, as is the case in a standard free recall task, the information score is uniform across serial positions.

Results and Discussions

We analyzed the narrative free recall and summarization datasets to assess temporal dynamics analogous to those of the standard free recall task.

Result 1: Summarization datasets demonstrate memory biases similar to those observed in free recall literature. The serial position curves for all datasets (Figures 3AEIM) showed enhanced recall or summary of sentences from the beginning of the source material, which aligns with the primacy effect typically observed in the standard free recall task (enhanced recall for items from the beginning of the list; Murdock, 1962; see also Figure 1B). Additionally, the Narrative free recall dataset and the RedditTIFU dataset showed enhanced recall or summary of sentences from the end of the source material (Figures 3AM), which aligns with the recency effect typically observed in standard free recall tasks (Murdock, 1962; see also Figure 1B). The probability of first recall curves for all datasets (Figures 3BFJN) showed subjects' tendency to initiate recall or summary from the beginning of the source material (primacy; see also Figure 1C) and, in the case of the Reddit TIFU dataset (Figure 3N), also from the end of the source material (recency; see also Figure 1C). Furthermore, Figures 3CGKO showed that when a recall or summary continues to unfold, it is more likely to transit to consecutive sentences (at smaller lags) in the source material than sentences that are farther away (at larger lags), which aligns with temporal contiguity effects (tendency to recall items from nearby positions in the study list; Kahana, 1996; see also Figure 1D) typically observed in the standard free recall task. The transition patterns also demonstrated more transitions at positive lags than negative lags, which aligns with the forward asymmetry effect (a bias towards making forward transitions; Kahana, 1996; see also Figure 1D) typically observed in standard free recall. To ensure that these results are not sensitive to our choice of similarity metric based on METEOR score during sentence alignment, we performed additional analyses using metrics BLEU (Papineni, Roukos, Ward, & Zhu, 2002) and cosine similarity (Pedregosa et al., 2011) on sentence embeddings (Reimers & Gurevych, 2019). These additional analyses yielded consistent conclusions and are not reported here due to space constraints. The observation that the summarization datasets demonstrated primacy, recency, temporal contiguity, and forward asymmetry effects, similar to those observed in free recall literature, supports our hypothesis that summarization involves processes of memory recall. The similarity between behavioral patterns in the Narrative free recall dataset (Figure 3, first row) and the three summarization datasets (Figure 3, remaining rows) provides further evidence for this hypothesis.

Result 2: The same set of memory biases remain after controlling for the structure of the source material. One might wonder if the observed primacy and recency ef-

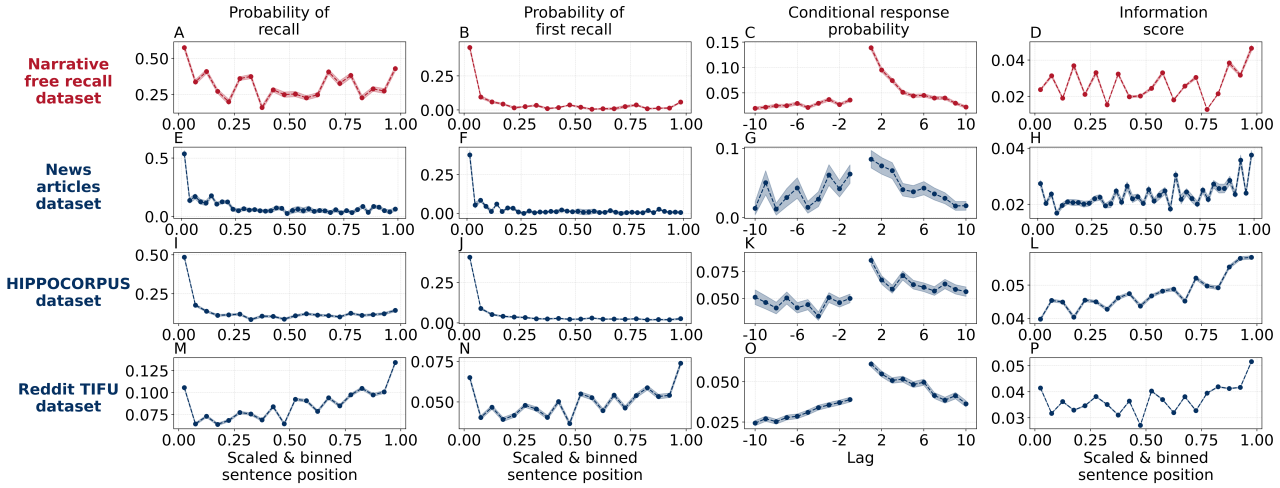

Figure 3: Behavioral patterns analyzed in the narrative free recall and summarization datasets. These patterns are the serial position curve, the probability of first recall, conditional response probability, and the average information score describing where important information is located in the source material for the (A–D) Narrative free recall, (E–H) News articles, (I–L) HIPPOCORPUS, and (M–P) Reddit TIFU datasets. The shaded error represents the standard error of the mean.

fects in the summarization patterns truly represent memory biases or are a result of the structure of the source material. For example, if sentences at the beginning of a naturalistic text tend to contain more important information, then one may consistently summarize information from the beginning of the text to improve the accuracy of summarization. To examine this possibility, we computed the average information score for each dataset, describing how important each sentence is in the source material. The information score plots (Figures 3DHLP) revealed that although primacy effects were observed consistently across all datasets, it is not always the case that important information is located at the beginning of the source material. In some cases, though, the primacy and recency effects did coincide with peaks in the information score – in the Narrative free recall dataset where recency was observed on the serial position curve (Figure 3A), the most critical information was on average concentrated at the end (Figure 3D), and for the News articles dataset where primacy was observed (Figure 3E), there was also a peak in information score observed at the beginning of the source (Figure 3H). For the Reddit TIFU dataset, where both primacy and recency were observed (Figure 3M), there were peaks in information score at both the start and the end of the source material (Figure 3P). For the cases where the probability of recall/summarization coincided with peaks in the information score, we analyzed whether the primacy and recency effects observed were contributed solely by where important information in the source material is. To do this, we filtered out trials that had important information at the start or end of the source material and reanalyzed the datasets for the same patterns. To eliminate the confound for primacy in the News articles and Reddit TIFU datasets, we excluded any trials that had an information score maxima in the first half of the source material. Similarly, to eliminate the confound for recency in the Narrative free recall and Reddit TIFU datasets, we excluded any trials with an information score maxima in the second half of the source. This filtering successfully eliminated the peak at the end of the average information score for the Narrative free recall (Figure 4D) and Reddit TIFU (Figure 4P) datasets and at the beginning for the News articles (Figure 4H) and Reddit TIFU (Figure 4L) datasets. After removal of respective trials, recency effects were still observed in the Narrative free recall (Figure 4A) and Reddit TIFU (Figures 4M–4N) datasets and primacy effects in the News articles (Figures 4E–4F) and Reddit TIFU (Figures 4I–4J) datasets. The persistence of these effects, after accounting for the structure of the source material, indicated that patterns observed in summarization are not solely driven by the location of important information but also reflect biases of memory recall. Even more interestingly, both the goal of summarizing important information and memory biases play a role in the final summarization behavior. This interaction could create tension and a non-intuitive behavior in some cases, as summarization tends to initiate from the beginning of the source material (Figure 3F) even when the most important information is concentrated towards the end (Figure 3H).

Result 3: Summarization datasets additionally demonstrate the list length effect similar to that observed in free recall literature. Finally, we examined whether patterns analogous to the effect of list length on recall initiation, as observed in standard free recall (longer presentation lists are associated with a reduced tendency to initiate recall at the first serial position; Ward et al., 2010) are present in the narrative free recall and summarization datasets. We divided trials into a short-length group and a long-length group using a median split based on the length of the source material (Figure

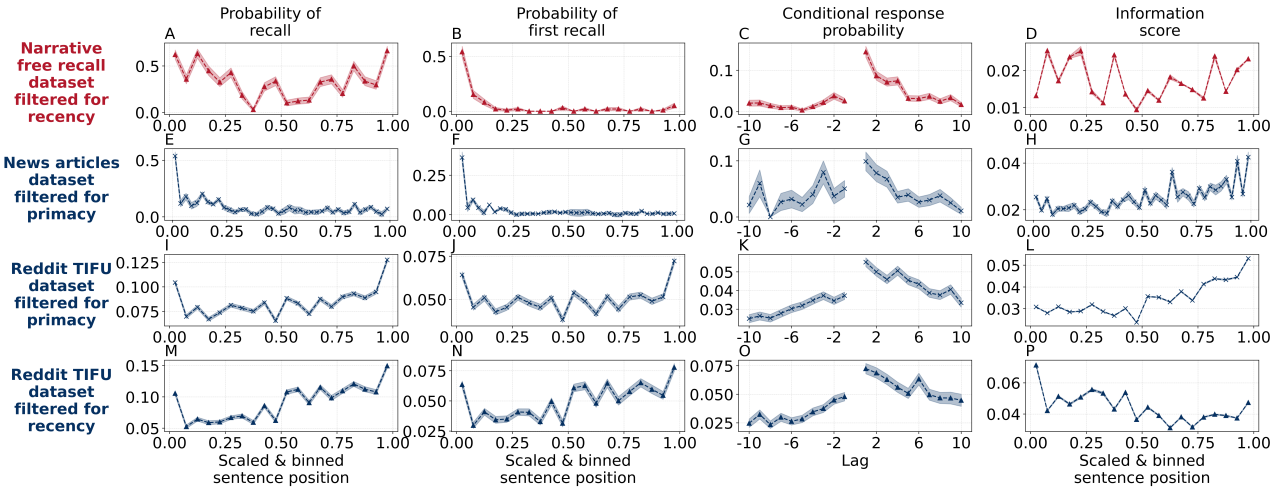

Figure 4: Behavioral patterns analyzed in the narrative free recall and summarization datasets, after filtering out a subset of trials based on information score. Similar effects of memory biases are observed compared to those before filtering (Figure 3) for the (A–D) Narrative free recall, (E–H) News articles, and (I–P) Reddit TIFU datasets. The shaded error represents the standard error of the mean.

5, third column), and plotted the probability of the first recall curve for each group (Figure 5, first column). We observed a higher proportion of trials to initiate summary from the first bin position in the short-length group in the summarization datasets, including the News articles (Figure 5E; $\chi^2(1, n_1 = 167, n_2 = 135) = 7.152, p = .007$), HIPPOCORPUS (Figure 5H; $\chi^2(1, n_1 = 1713, n_2 = 1066) = 13.337, p < .001$), and Reddit TIFU (Figure 5K; $\chi^2(1, n_1 = 21761, n_2 = 20378) = 504.608, p < .001$) datasets, consistent with that observed in standard free recall. However, we did not observe a significant effect of list length in the Narrative free recall dataset (Figure 5B; $\chi^2(1, n_1 = 449, n_2 = 394) = 1.400, p = .237$). The observed effect of list length on summary initiation, similar to that observed in free recall literature, provides additional evidence that summarization reflects processes of memory recall.

Conclusion

In this work, we analyzed the temporal dynamics of narrative free recall and summarization data and found that there is an alignment between the patterns seen in both tasks and standard free recall patterns. Summarization datasets demonstrate primacy, recency, temporal contiguity, and the effect of list length, similar to those observed in free recall literature. While our results support the hypothesis that summarization reflects characteristics of memory recall, a fruitful avenue for future work is to examine the differences between summarization and recall in addition to their similarities. As the objective of summarization is to retain key information from the source material, our results suggest that the way we summarize information is constrained by what we are able to recall from our memory. This opens up new opportunities in the future to use summarization as a novel task to study processes of memory recall in naturalistic settings involving complex and continuous stimuli. Laboratory studies of memory using naturalistic stimuli, such as movies or videos, can only collect a limited number of trials per experiment. On the other hand, many large datasets of human summarization are available in the computer science community, as human summaries are often used to evaluate machine learning models of summarization (Syed, Gaol, & Matsuo, 2021).

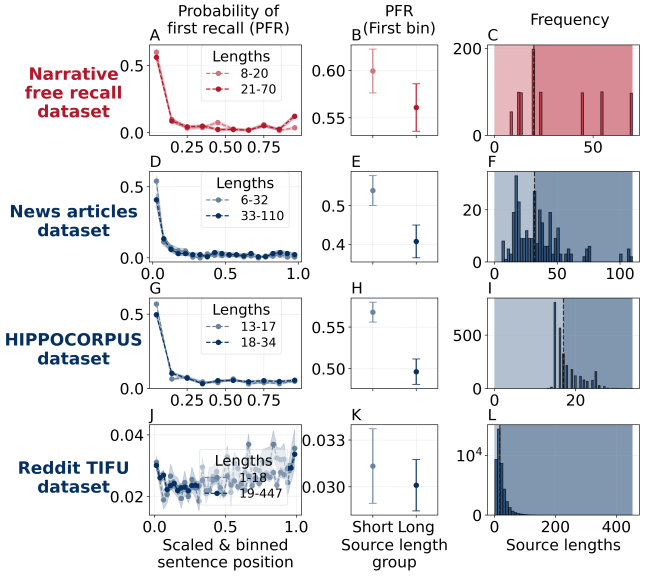

Figure 5: The probability of first recall for the short-length and long-length groups using a median split based on the length of the source material, analyzing the (A–C) Narrative free recall, (D–F) News articles, (G–I) HIPPOCORPUS, and (J–L) Reddit TIFU datasets.

References

Angne, H., Cornell, C. A., & Zhang, Q. (2024). A context-based model of collaborative inhibition during memory search. *Scientific Reports*, *14*(1), 27645.

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. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65–72).

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., Jin, S., & Zhang, Q. (2023). The role of episodic memory in storytelling: Comparing large language models with humans. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

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.

Coutanche, M. N., Koch, G. E., & Paulus, J. P. (2020). Influences on memory for naturalistic visual episodes: Sleep, familiarity, and traits differentially affect forms of recall. *Learning & Memory*, *27*(7), 284–291.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51–57.

Georgiou, A., Can, T., Katkov, M., & Tsodyks, M. (2023). Using large language models to study human memory for meaningful narratives. *bioRxiv*.

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.

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.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*(1), 103–109.

Kim, B., Kim, H., & Kim, G. (2019). Abstractive summarization of reddit posts with multi-level memory networks. In *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)* (pp. 2519–2531).

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Miller, J. F., Neufang, M., Solway, A., Brandt, A., Trippel, M., Mader, I., … others (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.

Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning* (pp. 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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1797–1807).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Rajaram, S., & Pereira-Pasarin, L. P. (2010). Collaborative memory: Cognitive research and theory. *Perspectives on Psychological Science*, *5*(6), 649–663.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992).

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.

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.

Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, *25*(2), 207–222.

Syed, A. A., Gaol, F. L., & Matsuo, T. (2021). A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access*, *9*, 13248–13265.

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.

Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining

the relationship between free recall and immediate serial recall: the effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1207–1241.

Weldon, M. S., Blair, C., & Huebsch, P. D. (2000). Group remembering: Does social loafing underlie collaborative inhibition? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1568–1577.

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of experimental psychology: General*, *130*(1), 29.

Zhang, Q., Griffiths, T. L., & Norman, K. A. (2023). Optimal policies for free recall. *Psychological Review*, *130*(4), 1104–1124.

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.