Human predictions of event sequences for novel videos

Caroline S. Lee (cl4353@columbia.edu)
Department of Psychology, Columbia University
New York, NY 10027, USA

Dana Lauren Villamen (dcv2109@barnard.edu)
Department of Psychology, Columbia University
New York, NY 10027, USA

Paulina Bertilsson (pmb2170@barnard.edu)
Department of Psychology, Columbia University
New York, NY 10027, USA

Mariam Aly (mariamaly@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94720, USA

Christopher Baldassano (c.baldassano@columbia.edu)
Department of Psychology, Columbia University
New York, NY 10027, USA

Abstract

Prior knowledge facilitates predictions in novel situations. Schematic templates for typical sequences of events, called event scripts, support event segmentation and linking of information across related contexts. Here, we present preliminary data comparing recall based on episodic memory and predictions based on event scripts. Participants viewed the first segment of instructional videos, some of which they had previously viewed, and then verbally recalled or predicted subsequent events. We used large language models (LLMs) to quantify the semantic content of the responses, enabling comparison between episodic memory and prediction-based knowledge.

Keywords: Prediction; episodic memory; LLMs; schema learning; event cognition; event segmentation

Introduction

Making predictions in novel scenarios requires us to draw on our past experiences in the world. And our prior knowledge can be organized into structured reusable templates, or schemas (Bartlett, 1995), that provide quick access to generalized information. A particular type of temporal schema, whereby a predictable set of ordered actions in time could be used to make sense of novel yet semantically related scenarios, is called a ”script” (Schank & Abelson, 2013). Scripts can be chunked into smaller, meaningful units called events (Bower, Black, & Turner, 1979), allowing for linking, or remembering, of related information underlying the formation of a broader temporal schema (Baldassano, Hasson, & Norman, 2018; Kurby & Zacks, 2008).

Evidence of segmenting temporally extended stimuli into events is present even for more complex audiovisual stimuli (movies, classroom material, stories, etc.) in the brain (Baldassano et al., 2017), and has been shown to elicit an anticipatory or predictive signal when the same stimuli are encountered repetitively (Lee, Aly, & Baldassano, 2021). However, previous work has not investigated the extent to which predictions for novel events, generated from schematized knowledge, differ from recollections drawn from our episodic memory. Here we examine these behavioral differences, using LLMs to compare and quantify human recalls and predictions of instructional videos, for which participants may use an event script from their prior knowledge or one learned (encoded) through the experiment.

Previous work has successfully used LLMs to quantify or predict human responses in various episodic memory tasks. Specifically, previous work has quantified the semantic similarity of recalls of varying lengths of narratives (Georgiou, Can, Katkov, & Tsodyks, 2025), as well as across different childhood experiences (Lee, Cohen, Hutchinson, Tottenham, & Baldassano, 2024). Pink et al., 2024 examined long term memory performance differences in sequential episodic memory tasks in humans compared to LLMs. However, LLMs have not yet been used to differentiate between recalls and predictions of novel scenarios.

Methods

Experiment

In a 2-part Prolific experiment, participants watched a set of instructional videos. Each video consisted of multiple steps, or events, culminating in a finished product (e.g. fixing a bike tire or making hummus). In part 1 of the experiment (encoding task), participants watched 8 - 16 instructional videos, which included all events. After each video, they were asked to place one animated GIF, depicting an event from the video, on a timeline indicating their estimate of when in the video it occurred (Figure 1A).

In part 2, participants watched the first event from the encoding task (part 1) as well as the first event from 8 - 16 new videos (Figure 1B). After each video, they were asked to verbally recall (old videos) or predict (new videos) subsequent events, relying on either their episodic memory or prior knowledge. Afterwards, participants completed the GIF-event placement task similar to part 1.

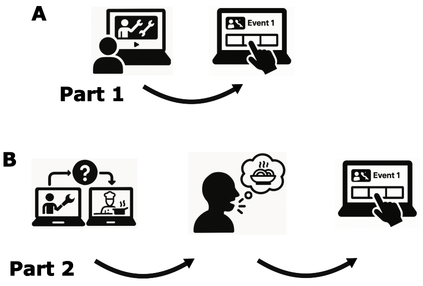

Figure 1: **A**. Participants viewed videos (left) in Part 1 of the experiment and placed a GIF of an event from the video on a timeline (right). **B**. In Part 2, participants watched the first event from both old and new videos (left), recalled or predicted the next events verbally (middle), and then completed the same event placement task as in Part 1 (right).

Stimulus set

Experimental stimuli were adapted from a computer vision dataset, COIN (Tang et al., 2019). Multiple events in 11,827 instructional YouTube videos of 180 various tasks, depicting instructions to complete a product or goal, were identified (durations in videos) and labeled (short text descriptions) by a computer vision model (Figure 2, top and middle), and used as guidelines to edit video stimuli. From this dataset, we selected 33 videos by the number of events (3 - 5 events) and for a variety of content (recipes, parts assembly, crafts, etc.).

Transcript processing

Each transcript was processed from audio to English text using Whisper, a pretrained speech-to-text model from Open AI (Radford et al., 2022) (Figure 2, bottom). Transcripts

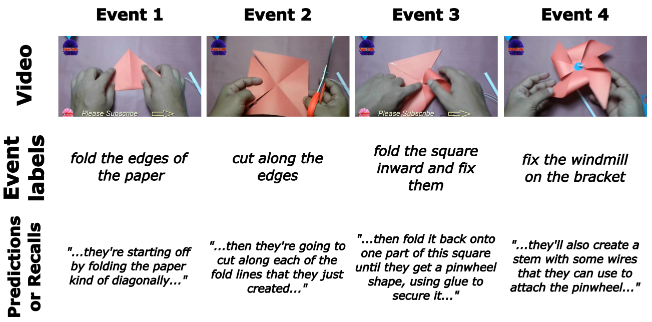
Event 1
Event 2
Event 3
Event 4
Video
Event labels
fold the edges of the paper
cut along the edges
fold the square inward and fix them
fix the windmill on the bracket
Predictions or Recalls
"...they're starting off by folding the paper kind of diagonally..."
"...then they're going to cut along each of the fold lines that they just created..."
"...then fold it back onto one part of this square until they get a pinwheel shape, using glue to secure it..."
"...they'll also create a stem with some wires that they can use to attach the pinwheel..."

Figure 2: Clips of events from an example instructional video (top) from the COIN dataset with corresponding event descriptions (middle). Participants predict or recall (example recall shown here) events from the videos (bottom).

were then corrected as needed to match corresponding audio recordings. Each transcript was scored for a) the number of events accurately reported (recalled or predicted) and b) the number of forward event transitions. Participants' event descriptions were matched for accuracy to the event description from the COIN dataset. The number of forward transitions were quantified by event descriptions proceeding further in time than the last (e.g. a description of event 3 *after* a description of event 1).

Sentence embeddings of text transcripts were generated using a pretrained sentence transformer model (RoBERTa) from Sentence BERT (Reimers & Gurevych, 2019). Embeddings were reduced to 2 dimensions for visualization with the UMAP algorithm (McInnes, Healy, & Melville, 2020). Similarities between pairs of transcripts (recall-to-recall, recall-to-prediction, prediction-to-prediction) within each video were computed with the Sentence BERT cross-encoder model.

Results

The accuracy of events reported differed by whether participants were remembering videos from the experiment or predicting unwatched events. The mean proportion of events accurately reported relative to the total possible number of events per video was significantly greater when participants were recalling from an earlier part of the experiment (i.e. relying on their episodic memory) (*M* = 0.87, *SD* = 0.16) than predicting subsequent events in videos that they had not yet encountered (i.e. relying on schematized prior knowledge) (*M* = 0.65, *SD* = 0.17); *t* = -10.90, $p < 0.001$ (Figure 3A, left). Similarly, the mean proportion of forward event transitions (out of all possible forward transitions) was also significantly greater when recalling videos from Part 1 (*M* = 0.85, *SD* = 0.16) than when predicting unwatched videos (*M* = 0.62, *SD* = 0.16); *t* = -11.62, $p < 0.001$ (Figure 3B).

Next, we examined the semantic similarity between individuals' predictions and recalls for each video. We found that, compared to similarities between random pairs of responses to a video, pairs of recalls for the same video tended to have significantly higher similarities ($p < 0.001$) while pairs of predictions were significantly less similar ($p = 0.008$) and recall-prediction pairs were the most dissimilar ($p < 0.001$) (Figure 3B, left), suggesting that there is more shared content in the recalls than the predictions. On examining sentence embeddings of descriptions of an example video, we can see that there is a dense cluster of recalls, and almost no overlap with a more sparsely distributed group of predictions (Figure 3B, right). This may be indicative of vastly differing spaces of prior knowledge from which participants can draw, even when prompted with the same first event.

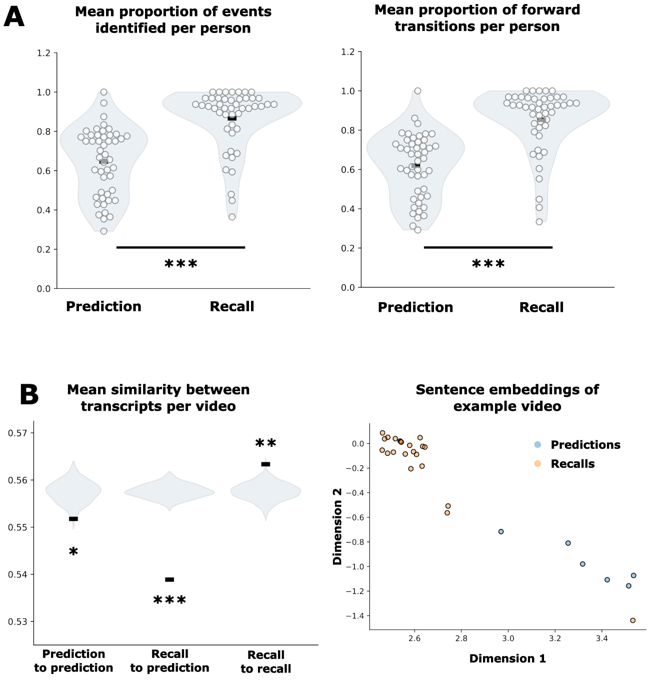

Figure 3: **A**. Both accuracy of events (left) and forward transitions (right) were higher on average on recalled videos from Part 1 than predicted events ($p < 0.001$). **B**. Pairwise comparisons of verbal descriptions of the same video were more similar than a null distribution when comparing 2 recalls ($p < 0.001$) and less similar for predictions ($p = 0.008$) and recall-to-predictions ($p < 0.001$) (left). Dimensionality reduction on sentence embeddings from a video here shows a dense cluster of recalls as compared to predictions.

Acknowledgments

Research reported in this study was supported by the National Eye Institute of the National Institutes of Health under award **NIH R01EY034436** to M.A. and C.B.

References

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709-721.

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, *38*(45), 9689–9699.

Bartlett, F. C. (1995). *Remembering: A study in experimental and social psychology*. Cambridge university press.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive psychology*, *11*(2), 177–220.

Georgiou, A., Can, T., Katkov, M., & Tsodyks, M. (2025). Large-scale study of human memory for meaningful narratives. *Learning & Memory*, *32*(2), a054043.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in cognitive sciences*, *12*(2), 72–79.

Lee, C. S., Aly, M., & Baldassano, C. (2021, April). Anticipation of temporally structured events in the brain. *Elife*, *10*.

Lee, C. S., Cohen, S. S., Hutchinson, S., Tottenham, N., & Baldassano, C. (2024). Neural and verbal responses to attachment-schema narratives differ based on past and current caregiving experiences. *bioRxiv*, 2024–09.

McInnes, L., Healy, J., & Melville, J. (2020). *Umap: Uniform manifold approximation and projection for dimension reduction.* Retrieved from `https://arxiv.org/abs/1802.03426`

Pink, M., Vo, V. A., Wu, Q., Mu, J., Turek, J. S., Hasson, U., ... Toneva, M. (2024). Assessing episodic memory in llms with sequence order recall tasks. *arXiv preprint arXiv:2410.08133*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision.* arXiv. Retrieved from `https://arxiv.org/abs/2212.04356` doi: 10.48550/ARXIV.2212.04356

Reimers, N., & Gurevych, I. (2019, 11). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing.* Association for Computational Linguistics. Retrieved from `https://arxiv.org/abs/1908.10084`

Schank, R. C., & Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press.

Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., ... Zhou, J. (2019). Coin: A large-scale dataset for comprehensive instructional video analysis. In *Ieee conference on computer vision and pattern recognition (cvpr).*