Behavior Research Methods (2025) 57:39
https://doi.org/10.3758/s13428-024-02569-z

ORIGINAL MANUSCRIPT

Large language models can segment narrative events similarly to humans

Sebastian Michelmann[1] · Manoj Kumar[2,3] · Kenneth A. Norman[2,3] · Mariya Toneva[4]

Accepted: 18 November 2024
© The Psychonomic Society, Inc. 2025

Abstract

Humans perceive discrete events such as "restaurant visits" and "train rides" in their continuous experience. One important prerequisite for studying human event perception is the ability of researchers to quantify when one event ends and another begins. Typically, this information is derived by aggregating behavioral annotations from several observers. Here, we present an alternative computational approach where event boundaries are derived using a large language model, GPT-3, instead of using human annotations. We demonstrate that GPT-3 can segment continuous narrative text into events. GPT-3-annotated events are significantly correlated with human event annotations. Furthermore, these GPT-derived annotations achieve a good approximation of the "consensus" solution (obtained by averaging across human annotations); the boundaries identified by GPT-3 are closer to the consensus, on average, than boundaries identified by individual human annotators. This finding suggests that GPT-3 provides a feasible solution for automated event annotations, and it demonstrates a further parallel between human cognition and prediction in large language models. In the future, GPT-3 may thereby help to elucidate the principles underlying human event perception.

Keywords Event boundaries · Event segmentation · Event cognition · Large language models

Introduction

Humans perceive events in continuous experience (e.g., "restaurant visits" and "train rides"; Zacks et al., 2007). Numerous studies have demonstrated the role of this inferred event structure in cognitive phenomena such as working memory (Jafarpour et al., 2022), time-perception (Bangert et al., 2020), and episodic memory (DuBrow & Davachi, 2013; Michelmann et al., 2023; Sargent et al., 2013). Furthermore, neural correlates of events have been identified in human functional magnetic resonance imaging (fMRI, Baldassano et al., 2017), electroencephalography (EEG, Silva et al., 2019) and intracranial EEG (iEEG, Michelmann et al., 2021). Moreover, a recent study identified "event cells" in rodents: single neurons that represented laps in a maze were invariant to lap-duration (Sun et al., 2020). Brunec et al. (2018) and Clewett et al. (2019) summarize in recent reviews how event boundaries shape behavior and human brain activity. Crucially, a substantial number of event-perception studies have focused on the comprehension of narrative text. In these studies, participants typically read a text (Zacks et al., 2009; Ezzyat & Davachi, 2011; Pettijohn & Radvansky, 2016), or listen to a spoken narrative (Zacks et al., 2009; Michelmann et al., 2021) with the instruction to segment the narrative into events. Importantly, individual raters agree substantially on the perception of event boundaries, drawing on an underlying shared perception of long-time-scale dynamics in narratives (Chen et al., 2017).

The ubiquitous influence of event structure on human cognition and neural activity suggests that there may be universal computational principles at work that make event segmentation useful for solving cognitive tasks (e.g., Lu et al., 2022). Zacks et al. (2007) and others (e.g., Shin and DuBrow, 2021) have posited that the segmentation of ongoing experience into discrete events is driven by prediction error. According

✉ Sebastian Michelmann
s.michelmann@nyu.edu

1 Department of Psychology, New York University, New York, NY, USA

2 Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA

3 Department of Psychology, Princeton University, Princeton, NJ, USA

4 Max Planck Institute for Software Systems, Saarbrücken, Germany

Published online: 03 January 2025

to this perspective, individuals generate predictions about the current event based on learned statistical information. For instance, we expect to be asked for our ticket on a train ride but would be surprised if the waiter at a restaurant asked us for our ticket instead of offering the menu. When these predictions begin to fail, it is inferred that the current event has ended and that a new event has begun, marking an event boundary. This hypothesized link between prediction and event segmentation suggests that large neural network models, which are likewise concerned with the prediction of upcoming information, may show sensitivity to event structure. In support of this view, recent studies have shown that measures of prediction error derived from large network models are significantly correlated with human annotations of event boundaries (Kumar et al., 2022; Mariola et al., 2022; Roseboom et al., 2022). Currently, however, no method has been designed to produce an automated event segmentation of text (see however, Wang et al. (2022) for automated detection of surprising event boundaries). While our method is currently agnostic to the underlying mechanisms that drive event segmentation, we note that it can explain substantial and significant variance in behavior (maximal correlation of $r = .372$) and that our segmentation performs within the expected range of individual human annotators (see Results).

Researchers have studied event cognition extensively in controlled settings where event structure is predetermined by the experimenter (e.g., DuBrow and Davachi, 2016), but recently there has been renewed interest in studying event cognition in more ecological settings (see Sonkusare et al., 2020). Studying event cognition with naturalistic stimuli like movies and stories typically involves laborious hand annotation of event boundaries that are often crowd-sourced from large behavioral samples in online experiments (e.g., Michelmann et al., 2021, 2023). Because annotations may vary between participants (Sasmita & Swallow, 2023), those large samples approximate a shared perception by averaging across many individual annotations. Notably, this approach is difficult to scale as the field becomes interested in studying how people process long naturalistic stimuli; getting suitably large numbers of participants to annotate boundaries in a long book would be a forbidding endeavor.

In this work, we propose to leverage recent developments in language models to enable progress in studying event cognition at naturalistic scales. Large language models that have been pre-trained to predict the next word over billions of text documents have swept up the natural language processing (NLP) field and led to state-of-the-art performance across NLP tasks. In parallel, cognitive neuroscientists have also begun evaluating these large language models with respect to their alignment with human behavior (Caucheteux et al., 2021; Golan et al., 2022; Kumar et al., 2022) and human brain activity during language comprehension (Jain & Huth, 2018; Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022), finding many significant correspondences. However, it is still an open question to what extent the inner workings of these large language models reflect human cognition.

The current approach examines whether a powerful large language model, GPT-3 (Brown et al., 2020), can segment events in a similar way to humans. Specifically, we assess its ability to perform event segmentation tasks using only zero-shot prompting (i.e., without any additional fine-tuning or in-context learning). We prompt GPT-3 with a similar (albeit simpler) prompt to what human participants are shown when asked to segment a naturalistic stimulus into meaningful events. Across a variety of stories, we find that GPT-3 is able to perform this task significantly above chance. Moreover, we demonstrate that GPT-derived event boundaries provide a good approximation of the consensus solution (obtained by averaging responses across people); the boundaries identified by GPT-3 are closer to the consensus, on average, than boundaries identified by individual human annotators. This suggests that GPT-3 segments event boundaries according to a shared human perception that is captured by averaging across large samples of human annotators. Finally, we derive a continuous measure of event probability from the log-probability of the newline character (that was used to indicate a new event) in the model output. This continuous measure of event boundary perception in the model significantly correlates with the continuous measure of agreement between human raters. This finding provides further evidence that GPT-3 can represent the event structure of a natural story in its predictions, rendering it a feasible tool for obtaining automated event annotations. We will make our code available to enable others to easily integrate the proposed approach into their analyses.

Methods

We develop a simple approach to enable GPT-3 to segment narrative stories into meaningful events, utilizing zero-shot prompting. We validate this approach across three stories of different lengths, and compare the resulting GPT-3 event segmentations with annotations from human participants.

Materials

Stimuli We tested our approach on narrative stories of three different lengths, which were previously used in psychological and neuroscientific experiments: 1) "Pieman", a short story of 7 min and 30 s that was presented at a live storytelling event at the Moth (O'Grady, 2008). "Pieman" is a humorous autobiographic description of the narrator's struggle with journalistic integrity when he started his career in journalism at the student newspaper of his university. 2) "Monkey in

the Middle", a 30-min-long audio podcast that appeared on "This American Life" (Chavis, 2017). It describes the events surrounding a famous legal case about copyright infringement; in this case PETA argued that the copyright to the photographs that monkeys took of themselves belongs to the monkeys. 3) "The Tunnel Under the World", a 25-min-long science fiction story that was broadcast on the radio show "X Minus One" (Pohl, 1956). The story follows Guy Burckhardt, who wakes up on June 15th with a vivid nightmare. As he goes about his day, he notices strange occurrences. The next day, he wakes up again on June 15th, realizing he is reliving the same day over and over.

Preprocessing Transcriptions of these auditory stories were used for their segmentation with GPT-3 in this study. Transcriptions were in plain text (UTF-8 encoding), no hyphens were present, and quotation marks and occurrences of ellipses were removed. Subsequently, the words were transformed into tokens; tokens are the units of text in large language models, for instance, one way of tokenizing the word "tokenizer " could be into "token" and "izer". The fast implementation of the pretrained GPT-2 tokenizer was used to tokenize the text (huggingface.co). The transcripts were aligned to the audio recordings via Penn Phonetics Lab Forced Aligner (Yuan & Liberman, 2008). This was done to later match behavioral button press responses (that had been recorded in milliseconds from the start of the audio) to positions in the text.

GPT-3 event segmentation

Generative Pre-trained Transformer 3 (GPT-3) is a large language model developed by OpenAI. It can generate human-like text based on the input it receives (Brown et al., 2020). GPT-3 is built on the transformer architecture (Vaswani et al., 2023), which uses self-attention mechanisms to process input text; it effectively weighs the importance of different words in a sentence to generate output. To learn how to generate its output, GPT-3 is trained on a vast corpus of text data from the Internet to predict the next word in a sentence. Therein, the model's weights are adjusted according to the error in its prediction.

Properties and hyperparameters The input that GPT-3 can process is limited to 4096 tokens. Only text within this so-called *context window* can affect predictions of the model. Additionally, a *temperature* parameter controls the degree to which randomness affects the model output. A lower temperature makes the output more deterministic, while a higher temperature makes it more creative. To increase reproducibility, the *temperature* parameter was set to 0. Additionally, only half of the width of the context window (i.e., $4096/2 = 2048$) was used, because the model output occasionally diverged from a word-for-word copy when longer context windows were used. The total context window was split between prompt and output tokens: The number of instruction tokens in the prompt was first deducted from the 2048 token window width, and the remaining tokens were then divided equally between tokens that were reserved for the story part of the prompt and the number of tokens that were requested in the output by setting the *max_tokens* parameter. For the story "Tunnel Under the World", this parameter setting proved to be too restrictive and occasionally led to truncated output; to address this, additional padding of 512 tokens was added to the *max_tokens* parameter for the "Tunnel" story.

Prompting In order to prompt GPT-3 to segment a continuous story into events, a completion was created with the model version "text-davinci-002" (beta.openai.com/docs/models/). In completion, the model is asked to complete the provided prompt by predicting upcoming tokens. The specific prompt for the model was: "*An event is an ongoing coherent situation. The following story needs to be copied and segmented into events. Copy the following story word-for-word and start a new line whenever one event ends and another begins. This is the story:*". After this instruction, a segment from the story was inserted (we split the full story into shorter segments, see next paragraph). Finally, after a new line, an additional prompt segment was added to refresh the instruction: "*This is a word-for-word copy of the same story that is segmented into events:*". In a variation of the instruction, the qualifier "*long*" was added before every occurrence of the word event in the prompt, with the exception of the first definition. This was done to obtain a coarse-grained event segmentation.

Sliding window for long stories The context length of the GPT-3 model that we used was not long enough to encompass each of the three stories. Therefore, we used a sliding window approach to split the stories into segments that fit into the context length of GPT-3. The length of a story segment was determined based on the number of input and output tokens that fit into the context window (see below for some notes on the width that was used for the context window). The output text that the model generated was then split into events at newline characters. After processing the first story segment this way, we defined the next segment by starting at the final model-identified event boundary in the first segment, and then continuing forward with as much of the story as would fit given constraints on context window size. We continued with this process until the entire story had been fed into the model. Note that, in this scheme, any "trailing" tokens falling between the final model-identified event boundary of a segment and actual end of that segment were simply reassigned to be the start of the next segment. I.e., when the final event in a story segment stopped before the end of the full story, the next story segment was taken from the beginning of the final

event of the previous segment and continued into the story for the next completion request. The segmentation was complete after all text from a given story had been parsed. Note that our sliding window does not produce overlap in event segmentation, because all information in the final event (where no further event boundaries are detected) becomes the beginning of the next segment, i.e., there is no overlapping event boundary and the event boundaries in the subsequent segment can simply be concatenated to the event boundaries in the previous segment.

Continuous segmentation probability Another way to analyze behavioral responses is to compute the probability of button presses at every moment in the story. Similarly, predictions in GPT-3 can be treated as probabilistic. In order to compare the probability of event boundaries directly, we, therefore, log-transformed human event annotations to obtain the log-probability of a button press throughout the story. For GPT-3, we derived a continuous output variable from the log-probability of the newline character for every token in the story (averaged across 6 iterations of the model) with the following procedure: The log-probability of the top 5 predicted tokens was requested in the completion (which was the maximum number of token-probabilities that could be requested by the user at the time of preparation of this manuscript). The log-probability of the newline character was then used to assess to what degree the model predicted an event boundary at any given token in the story. Whenever the newline character was not among the top 5 predicted tokens, the log-probability was not defined. Note that, on average, 57.375 % of tokens entailed the newline character among the top 5 predictions (min = 52.833 %, max = 61.438 %). To interpolate the token-probabilities onto the same time axis as the behavioral data, the onset and offset times of words in the transcript were used. Words from the transcript were converted to tokens with the pretrained GPT-2 tokenizer; whenever a word was converted to more than one token, the interval spanned by its onset and offset was evenly split among the resulting tokens which yielded token-onset and -offset times. Next, all words from a given GPT-3 generated response were tokenized and word embeddings from the GPT-2 Model transformer (GPT2LMHeadModel) were used to map model-output tokens onto the list of aligned tokens from the transcript via dynamic-time-warping (using the fastdtw implementation Salvador and Chan, 2007, distance metric: Euclidean distance). The warp path was then applied to the token-onset and -offset times, which resulted in an onset and offset for every token that was generated by GPT-3. Finally, a vector of NaN values of the same length as the time-axis of the story was generated; for those tokens where a newline character was among the top-5 predicted tokens, the value of that vector was set to the log-probability of the newline character in the whole interval between the token's onset and offset time. The outcome of this transformation was a vector of the same length as the behavioral response vector from the human participants; the vector held the log-probability of the newline character whenever it was among the top-5 predicted tokens and NaN otherwise. NaN values were then interpolated with one-dimensional linear interpolation.

Human event segmentation

To assess the quality of GPT-3's automated event segmentation, we compared the event boundaries segmented by GPT-3 to those annotated by large numbers of human participants. Data from the human behavioral event segmentation tasks were previously analyzed and published, and were made available by the authors upon our request. In this subsection, we describe the details of these datasets, which are summarized in Table 1.

Experimental paradigm Participants in the first two studies were recruited via Amazon's Mechanical Turk (Buhrmester et al., 2011). Event segmentations for the "Pieman" story were obtained from 205 participants that listened to the story

Table 1 Word count and number of behavioral events for each of the three stories

Story	N words	N events (consensus)	N events (individual button presses)
"Pieman (run 1)"	1137	19	26.41 (SD = 17.71, range = 3–86)
"Pieman (run 2)"		20	24.48 (SD = 16.16, range = 3–72)
"Monkey"	6028	43	
"Monkey (Q1)"			44.03 (SD = 28.92, range = 3–107)
"Monkey (Q2)"			29.97 (SD = 20.88, range = 3–77))
"Monkey (Q3)"			20.61 (SD = 14.86, range = 2–68)
"Monkey (Q4)"			27.05 (SD = 22.14, range = 1–88)
"Tunnel"	4418	30	72.5 (SD = 33.99, range = 15–133)

Note that annotations for "Pieman" were collected twice, and that "Monkey in the Middle" was collected across four sub-groups that annotated a quarter of the story each

twice and also segmented the story twice (Michelmann et al., 2021), i.e., data from two behavioral runs were available. The verbatim instruction for these participants was: "press the space bar every time when, in your judgment, one natural and meaningful unit ends and another begins". Event segmentation for "Monkey in the Middle" was obtained with the same instructions, however, batches of participants annotated 4 sections of the story that were similar in length: All participants listened to the full story but only segmented their respective section; a total of 30 participants segmented the first quarter, 35 participants segmented the second quarter, 31 participants segmented the third quarter, and 42 participants segmented the final quarter of the story (Kumar et al., 2022). Event segmentation for "Tunnel Under the World" was obtained from 10 participants that participated in person (Lositsky et al., 2016).

Human consensus estimation, event annotation, and sentence boundary alignment Aggregated event boundaries were derived for all data by computing smoothed averages across participants and detecting local peaks in the time course of the agreement. Specifically, for the "Pieman" story, individual participants' data were aggregated in response vectors at a millisecond resolution that were set to 1 if a response was recorded within the surrounding second (i.e., within 500 ms in either direction) and to 0 otherwise. For the stories "Monkey in the Middle" and "The Tunnel Under the World", data were aggregated in response vectors in the same way; however, because less data were available, a response was marked within the 2-s interval around button presses (i.e., within 1 s in either direction).

The averaged response vector was then smoothed with a Gaussian kernel and thresholded. For the "Pieman" story time-courses were smoothed with a Gaussian kernel of 2-s width; time-courses of "Monkey in the Middle" and "The Tunnel Under the World" were smoothed with a kernel with a standard deviation of 0.2 s, which corresponds to an interval of approximately 1 s in the smoothing kernel. Finally, the smoothed agreement was thresholded at the 95th percentile for the "Pieman" story. For the stories "Monkey in the Middle" and "The Tunnel Under the World", boundary points were computed using the Python function `scipy.signal.find_peaks` and thresholded at the 90th percentile.

Within clusters of neighboring points, each cluster's maximum was then taken as a local peak that indicated an event boundary. For "Monkey in the Middle" and "The Tunnel Under the World" some duplicate boundaries were subsequently removed manually if they appeared after the same clause (the smaller boundary was discarded). Note that these local peaks represent the "consensus solution" of event boundaries. In the "Pieman" story 19 event boundaries were detected on the first and 20 event boundaries were detected on the second behavioral run. For the story "Monkey in the Middle" 43 event boundaries were detected, and for "Tunnel Under the World" 30 event boundaries were found (see Table 1). Crucially, all boundary detection was done prior to this work in the course of the original studies described in Michelmann et al. (2021) and Kumar et al. (2022).

For the purposes of this study, we attributed behavioral event boundaries (individual button presses or boundaries from aggregate responses) to the nearest sentence boundary. The distance to each sentence boundary was computed in milliseconds, where the time of each sentence boundary was based on the average between the offset of the last word of the previous sentence and the onset of the first word of the next sentence. For the "Pieman" story, event boundaries were adjusted on average 1.068 s ($min = 0.018$, $max = 6.209$) for the first, and 0.792 s ($min = 0.004$, $max = 5.594$) for the second behavioral run to correspond with sentence boundaries. For "Monkey in the Middle" the average adjustment was 1.64 s ($min = 0.012$, $max = 7.568$), and boundaries in "Tunnel Under the World" were adjusted by an average of 3.368 s ($min = 0.076$, $max = 27.981$). Regarding the number of event boundaries, on the individual subject level, participants marked on average 26.41 sentence boundaries in the "Pieman" story as event boundaries on the first behavioral run ($SD = 17.71, min = 3, max = 86$) and found 24.48 boundaries on the second run ($SD = 16.16, min = 3, max = 72$). For "Monkey in the Middle" the average numbers of event boundaries in the four quarters were 44.03 ($SD = 28.92, min = 3, max = 107$), 29.97 ($SD = 20.88, min = 3, max = 77$), 20.61 ($SD = 14.86, min = 2, max = 68$), and 27.05 ($SD = 22.14, min = 1, max = 88$), in "Tunnel Under the World", individual participants marked on average 72.5 sentence boundaries as event boundaries ($SD = 33.99, min = 15, max = 133$).

Attributing button presses to sentence boundaries made it possible to directly compute a distance metric between human-annotated and GPT-3-derived event segmentation that always coincided with sentence boundaries. Specifically, we computed the Hamming distance between the human and model-based boundary vectors. To additionally obtain a continuous measure of boundary detection in humans, the response vectors from individual participants were averaged (without smoothing). Since the resulting ratio of participants that responded within a given second can be interpreted as a response probability, the log of that ratio was computed for comparison with the model's log-probability output.

Statistical comparison of event annotations

To compare the distances between event annotations at the sentence level, a vector was created that held zero for each sentence boundary that was not an event boundary, and one

for each sentence boundary that was also an event boundary. Two event annotations were then compared by computing the Hamming distance between their respective vectors (i.e., the proportion of disagreeing components). Hamming distance typically represents the number of substitutions required to change one vector into another (Hamming, 1950); here we use a normalized version that computes the proportion of substitutions (e.g., vectors [1, 0, 0] and [1, 1, 0] have a Hamming distance of 1/3 = 0.333 because only one bit out of three has to be changed to make the vectors identical). This measure has the advantage that it is sensitive to incorrect boundary placement and incorrect omissions of boundaries, which both increase the distance; furthermore, it is invariant to the number of sentences in a story. Subsequently, the true Hamming distance was compared to the distribution of distances under 100, 000 random sentence permutations and a p value was obtained as the proportion of distances that were smaller under random conditions. To compare the similarity between individual participants' segmentations and the consensus solution with the similarity between GPT-3-derived segmentations and the consensus solution, the Hamming distance between every participant's sentence-level segmentation and the sentence-level segmentation from the consensus solution was computed. Likewise, the Hamming distance between each GPT-derived solution and the consensus solution was computed. Finally, these Hamming distances were compared between GPT-3 and the human participants with an independent sample two-sided t-test. To compare the log-probability between the behavioral event annotations and the newline characters predicted by GPT-3, we first interpolated NaNs in the vector that held the model output. In the human data, the button-press probability was already defined at each time-point and could be log-transformed for comparison. However, there were some moments where no participant had pressed a button, resulting in time-points where the log-transformation of the probability is undefined. To handle these missing values, we therefore interpolated the log-probability of button press at those time points. Subsequently, the cross-correlation between the vectors was computed and the zero-lag correlation was tested statistically against non-correlation.

Results

Event segmentations from GPT-3 significantly align with consensus human annotations

We segmented the text from the stories "Pieman", "Monkey in the Middle" and "Tunnel Under the World" into discrete events using GPT-3 (version "text-davinci-002"). "Pieman" was segmented into 23 events by GPT-3 (compared to 19 events in the averaged human annotations based on the first behavioral run and 20 events in the annotations from the second behavioral run, see Fig. 1 for an example comparison of GPT-3-derived event boundaries and human button presses). For "Monkey in the Middle" we obtained 88 events from the model output (compared to 43 events in the averaged behavioral sample) and for "Tunnel Under the World", we obtained 139 events (vs. 30 behavioral events). Next, we computed the Hamming distance between the boundary-vectors that were derived from the GPT-3 model output and the boundary-vectors of annotations from the average human response. For "Pieman" we obtained a Hamming distance of 0.255, [95% CI: 0.234 0.277], when comparing to the human boundaries from the first behavioral run and a Hamming distance of 0.245, [95% CI: 0.223 0.266], when comparing to event boundaries from the second behavioral run. These distances were significantly smaller than what was obtained from comparisons under random permutation of sentence order ($p = 0.024$, and $p = 0.009$, respectively). For "Monkey in the Middle" we obtained a Hamming distance of 0.25 ($p = 0.044$), [95% CI: 0.243 0.257], and "Tunnel Under the World" yielded a Hamming distance of 0.264 ($p < 0.001$), [95% CI: 0.26 0.268].

The model found more event boundaries than the number of behavioral event boundaries in all stories. However, we also observed a wide range of individual human boundary annotations (see Section "Methods"). We next set out to decrease the number of event boundaries in the model output by prompting the model to segment the story into *long* events.

Segmenting *long* events We observed that prompting GPT-3 to segment *long* events results in segmentation that is even closer to the human consensus. Prompting the model to segment into *long* events resulted in 14 events in the "Pieman" story (vs. 23 without the *long* qualifier), 59 events in "Monkey in the Middle" (vs. previously 88 without the *long* qualifier) and 76 events in "Tunnel Under the World" (vs. 139). When comparing each model-derived event segmentation to human annotations, we obtained a Hamming distance of 0.223, [95% CI: 0.213 0.234], when comparing to human annotations from the first "Pieman" run and 0.191, [95% CI: 0.181 0.202] when comparing to human annotations from the second "Pieman" run. These distances exceeded the distances that were observed under sentence level shuffle with a statistical trend compared to the first behavioral run ($p = 0.069$) and with statistical significance when compared to the second behavioral run ($p = 0.004$). For the story "Monkey in the Middle" we obtained a Hamming distance of 0.193 ($p = 0.024$), [95% CI: 0.186 0.2], and for "Tunnel Under the World" the Hamming distance between human annotations and the segmentation based on GPT-3 was 0.168 ($p = 0.002$), [95% CI: 0.164 0.172].

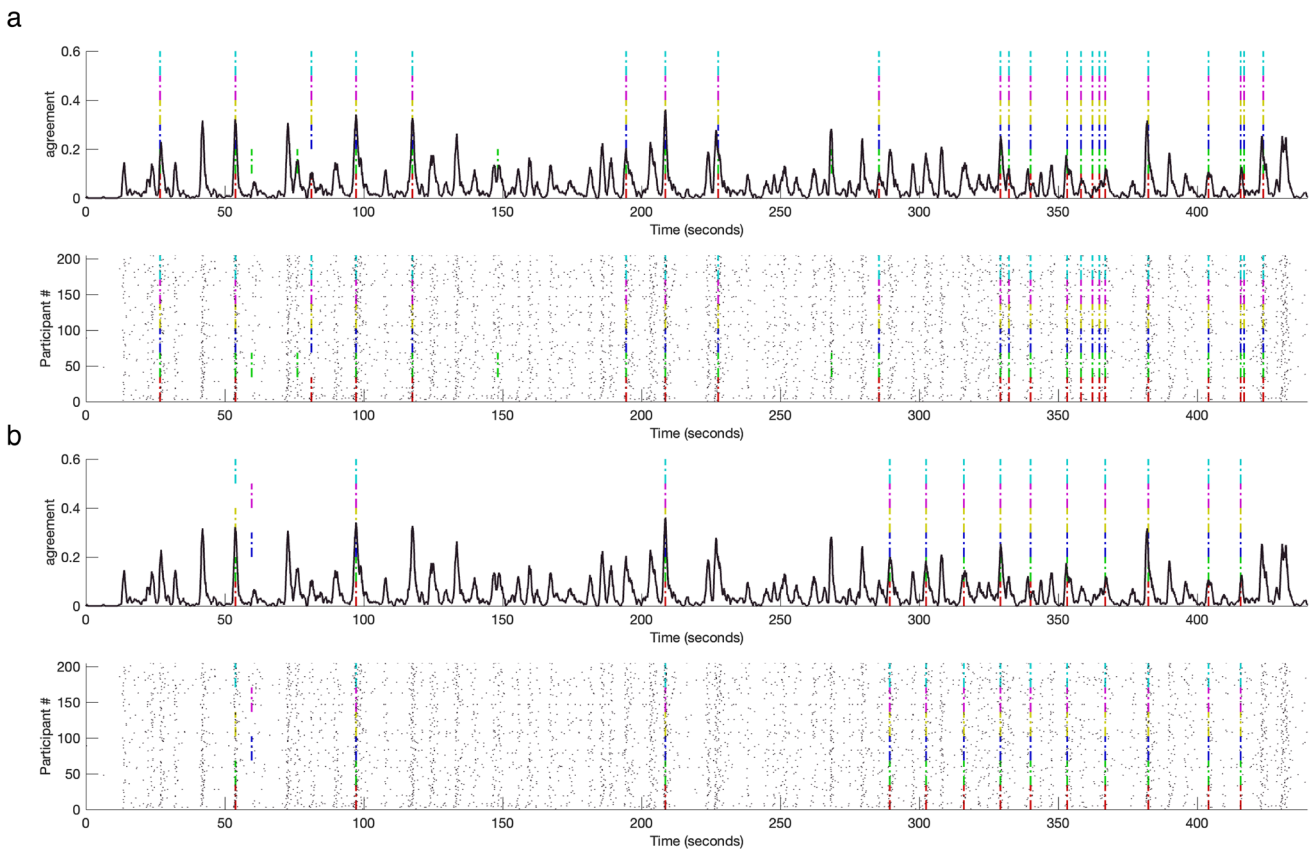


Fig. 1 Example of human event boundary annotations and GPT-3-derived event boundaries. *Vertical bars* show event boundaries detected by GPT-3 without the *long* qualifier (**a**) and with the *long* qualifier (**b**); *colors in vertical bars* denote different replications. The overlaid time-course of agreement (*top*, **a** and **b**) is from the second behavioral run of the "Pieman" story. The overlaid raster plot (*bottom*, **a** and **b**) shows individual button presses of each participant. Note that for boundary detection, the data were additionally smoothed with a Gaussian kernel; however, displayed time-courses reflect the non-smoothed agreement

Reproducibility. Because GPT-3 does not produce fully deterministic results (even with a *temperature* setting of 0), we next wanted to replicate our findings by repeating the model-based event segmentation five additional times for each story (see Fig. 1). With the "Pieman" story, we observed only one diverging replication where the model found 26 events that had a Hamming distance of 0.287 [95% CI: 0.266 0.309] to the first ($p = 0.056$) and 0.255 [95% CI: 0.234 0.277] to the second behavioral segmentation ($p = 0.006$) under the original event segmentation prompt. With "Monkey in the Middle" we observed two additional solutions that both had a Hamming distance of 0.25 [95% CI: 0.243 0.252] to the behavioral data with 86 events ($p = 0.074$) and 85 events ($p = 0.069$), "Tunnel Under the World" also yielded two additional solutions of 140 and 139 events that had Hamming distances of 0.266 [95% CI: 0.262 0.27] and 0.2622 [95% CI: 0.258 0.266] to the human annotations (all $ps < 0.001$). Note that the different solutions based on GPT-3 were very similar to each other with maximum Hamming distances between different solutions of 0.053 for "Pieman", 0.042 for "Monkey in the Middle", and 0.008 for "Tunnel Under the world". When prompting the model to segment the stories into *long* events, we also found diverging solutions in the five replications. For "Pieman" one additional solution yielded 14 events and had a Hamming distance of 0.245, [95% CI: 0.234 0.255], to human annotations from the first behavioral run ($p = 0.216$, n.s.) and 0.213, [95% CI: 0.202 0.223], to boundaries from the second run ($p = 0.022$). For "Monkey in the Middle" one additional solution with 56 events had a Hamming distance of 0.186 ($p = 0.016$), [95% CI: 0.178 0.193], to the behavioral data. For "Tunnel Under the World" two additional solutions with 65 and 68 events were found that had Hamming distances of 0.145, [95% CI: 0.141 0.15], and 0.152, [95% CI: 0.148 0.156], to the behavioral data ($ps < 0.001$). Again, different solutions based on GPT-3 were very similar to each other with maximum Hamming distances between different solutions of 0.021 for "Pieman", 0.007 for "Monkey in the Middle", and 0.098 for "Tunnel Under the World".

Comparison to other normed measures of similarity

Previous work on event segmentation in humans has reported normalized measures of correlation between raters. Specifically, Kurby and Zacks (2011) report a normalized point-biserial correlation, where correlations are expressed relative to the minimum and maximum correlation possible given the number of boundaries a participant identified. We adopted this approach for comparison purposes and expressed similarity values (1-Hamming distance) relative to the maximum and minimum similarity, given the number of boundaries that GPT-3 and human annotators identified. Without the *long* qualifier, we obtained relative similarity values of 0.444 and 0.482 for the first and second behavioral run of the "Pieman" story, averaging across all replications. For "Monkey in the Middle", the average relative similarity was 0.321 and for "Tunnel Under the World" we obtained a relative similarity of 0.655. When including the *long* qualifier, we obtained an average relative similarity of 0.359 and 0.513 with the first and second behavioral run of the "Pieman" story. We further observed a relative similarity of 0.262 for "Monkey in the Middle", and 0.379 for "Tunnel Under the World". Note that all p values remained identical because the permutation statistic is scale invariant.

GPT-3 event segmentation is statistically closer to the human consensus than individual humans

The significant similarities between GPT-3-derived event annotations and the event segmentation from human annotations suggest that GPT-3's segmentation reflects a shared human perception. Importantly, human-derived event boundaries are also a noisy approximation of an underlying shared event structure. Consequently, the question arises of how GPT-3 compares to typical human annotations that can be recorded in online experiments. In the "Pieman" story, for instance, GPT-3-derived event boundaries were consistently more similar to event boundaries from the second behavioral run compared to the first. In previous work, Michelmann et al. (2021) demonstrated that the second behavioral run is characterized by more agreement between participants, faster responses, and stronger correlations with neural measures of interest in electrocorticographic recordings from patients that listen to the story "Pieman". That is, the second behavioral run of "Pieman" may represent a better approximation of a perceived event structure. To test how GPT-3-derived event boundaries compare to individual human responses in measuring an underlying "consensus solution", we compared the average Hamming distance between GPT-3-derived solutions and the aggregate behavioral response (on the one hand) to the average Hamming distance between individual human participants' responses and the aggregate behavioral response (on the other hand). That is, we compared the similarity to the consensus solution between GPT-3 and individual human participants. We first tested this for the original event segmentation prompt (without the qualifier *long*) for those stories that had enough annotations to derive a meaningful aggregate response ("Pieman": N = 205, and "Monkey in the Middle": N >= 30 per quarter, see Fig. 2). For "Pieman" we found that individual participants' segmentations had an average Hamming distance of 0.281 ($SD = 0.116$) to the consensus solution on the first behavioral run. The average Hamming distance of GPT-3-derived segmentations to the consensus solution was significantly lower (0.261; $SD = 0.012$, $p = 0.045$). On the second behavioral run, individual participants' segmentations had an average Hamming distance of 0.252 ($SD = 0.099$) to the consensus solution; the GPT-3-derived segmentations still had a lower average Hamming distance (0.246, $SD = 0.004$), however, the difference was no longer significant ($p = 0.46$). Putting these results together: On the first run, GPT-3 does a better job, on average, of approximating the consensus than individual human participants; on the second run, human participants increase their segmentation performance, so the

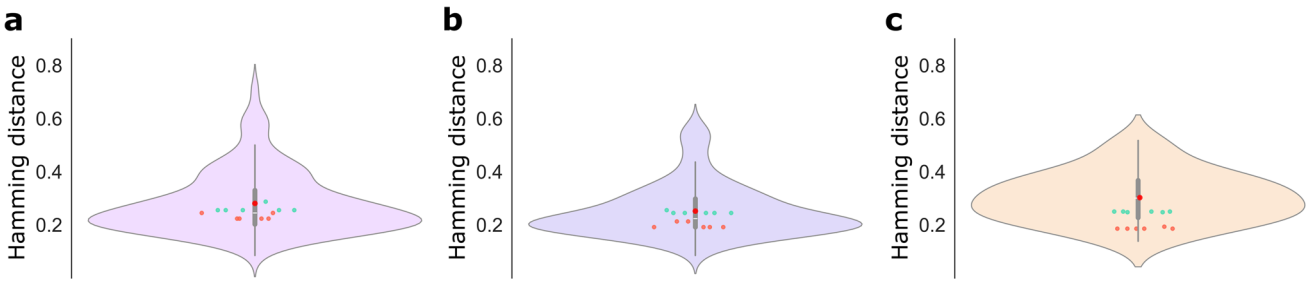


Fig. 2 Distribution of distance to the consensus solution for human data and GPT-derived segmentation. *Violin plots* show the distribution of distances to the consensus for individual human button presses in the first (**a**) and second (**b**) behavioral run of the "Pieman" story, and for the story "Monkey in the Middle" (**c**). The *thick gray bar* represents the interquartile range, the *red dot* is the mean, the *white line* is the median, and the *thin gray line* displays the extremes of the distribution excluding outliers. *Colored dots* represent the distance of the GPT-3-derived solution to the consensus, with (*orange*) and without (*green*) the *long* qualifier

GPT-3 segmentation is no longer significantly better. For "Monkey in the Middle", we treated concatenated response vectors from the four different segments (selected at random without replacement) as a single participant. These 30 combined data sets had an average Hamming distance of 0.302 ($SD = 0.094$) to the consensus solution (derived from the average across all participants). Again, the GPT-3-derived solutions were significantly more similar to the consensus solution with an average Hamming distance of 0.249 ($SD = 0.001$, $p = 0.005$) supporting the idea that GPT-3-derived segmentations are closer to a consensus solution that large samples of human raters converge on (see Fig. 2).

We next repeated this analysis with the *long* event segmentation prompt. For "Pieman" the average Hamming distance of *long* GPT-3-derived segmentations to the consensus solution was 0.23 ($SD = 0.01$) for the first behavioral run and 0.199, ($SD = 0.01$) for the second behavioral run. Both of these values were significantly lower than the average for human annotations ($p < 0.001$ in both cases). For "Monkey in the Middle", the average Hamming distance of *long* GPT-3-derived segmentations to the consensus solution was 0.187 ($SD = 0.003$), which was significantly lower than the average for human annotations ($p < 0.001$). The lower Hamming distances under the *long* event segmentation prompt suggest that these long events may be closer to shared human perception.

Comparison of boundary probabilities between GPT-3 and human annotators

The event boundaries from human participants are those moments where a substantial amount of people pressed a button within a certain time window (i.e., peaks in agreement). That is, human event boundaries are discrete measures derived from a continuous variable (ratio of responses). Another way to treat this measure, however, is as the probability of button presses at every moment in the story. Similarly, predictions in GPT-3 can be treated as probabilistic. In order to compare the probability of event boundaries directly, we, therefore, log-transformed human event annotations to obtain the log-probability of a button press throughout the story. For GPT-3, we derived a continuous output variable from the log-probability of the newline character for every token in the story (averaged across all 6 iterations of the model) and we interpolated those values onto the same time axis (in milliseconds, based on the audio; see Section "Methods" for detail). We then computed the cross-correlation between those time courses (note that we did not expect a specific lag between the time-courses, however, because human responses may entail a perception-to-action delay, we computed the cross-correlation to obtain an upper bound of the true correlation, see Fig. 3). With the original prompt, we obtained a zero-lag correlation of $r = 0.308$ ($p < 0.001$, maximum correlation of $r = 0.32$ at lag $303ms$) for the "Pieman" story's

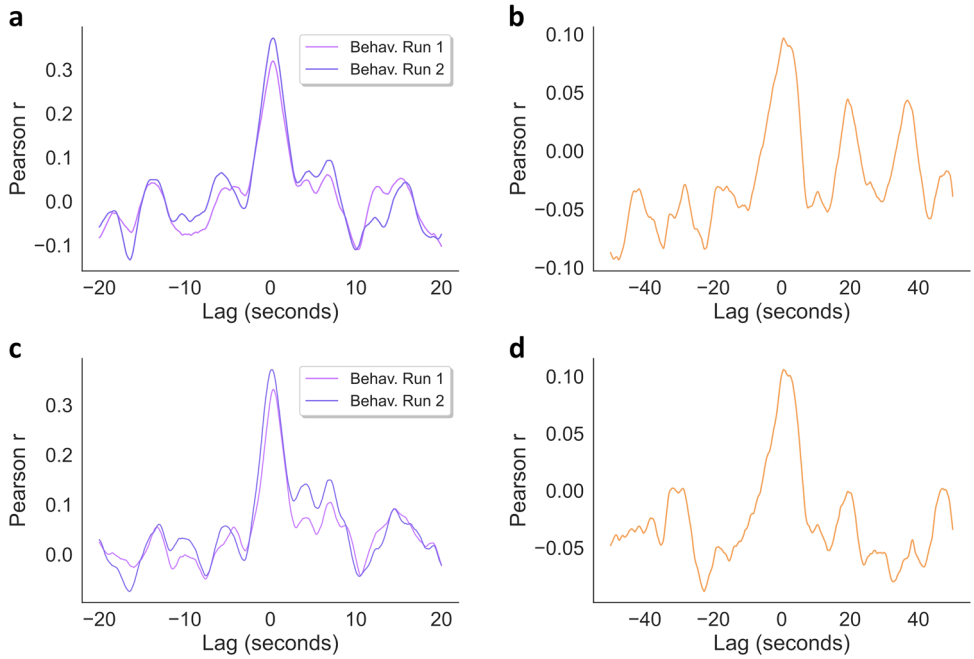


Fig. 3 Cross-correlations between log-probability of button press in the behavioral samples and the log-probability of the newline character in the model output (**a**) "Pieman" original prompt (**b**) "Monkey in the Middle" original prompt (**c**) "Pieman" *long* event prompt (**d**) "Monkey in the Middle" *long* event prompt

first behavioral run and a zero-lag correlation of $r = 0.362$ ($p < 0.001$, maximum correlation of $r = 0.372$ at lag $330ms$) for the second behavioral run. For "Monkey in the Middle" we obtained a zero-lag correlation of $r = 0.092$ ($p < 0.001$, maximum correlation of $r = 0.097$ at lag $553ms$). Under the *long* event prompt, we obtained a zero-lag correlation of $r = 0.311$ ($p < 0.001$, maximum correlation of $r = 0.331$ at lag $362ms$) with the first behavioral run of the "Pieman" story and a zero-lag correlation of $r = 0.369$ ($p < 0.001$, maximum correlation of $r = 0.371$ at lag $169ms$) with the second behavioral run. For "Monkey in the Middle" we observed a zero-lag correlation of $r = 0.102$ ($p < 0.001$, maximum correlation of $r = 0.106$ at a lag of $552ms$). Note that the lower correlation for the "Monkey in the Middle" story likely stems from a noisier estimate of the moment-by-moment time course of button-press probability due to the smaller sample size. When comparing discretized boundaries, similar Hamming distances are observed (compare Table 2). Taken together these findings are evidence that GPT-3 can represent the event structure of a naturalistic story in its predictions, rendering it a feasible tool for obtaining event annotations.

Discussion

Here we demonstrate that GPT-3 can segment a continuous narrative into events, akin to the way human annotators perform this task. A key benefit of our approach is that it is more scalable to very long naturalistic stimuli as obtaining human-derived annotations for very long stimuli can be prohibitively labor- and resource-intensive. GPT-3 derived boundaries were also significantly closer to the "consensus" solution, on average, than boundaries identified by individual human annotators. These findings suggest that GPT-3 is a feasible method for cognitive scientists that are interested in event segmentation. When individual human annotations are expected to be noisy and large behavioral samples are not available, a language model that has been trained on 45TB of text may be a useful alternative to obtain normative information about human text perception.

Automated methods of event segmentation have the potential to propel the field towards the study of human cognition at increasingly larger time scales. For instance, our approach can enable annotations of a whole book, which has been a forbidding endeavor so far. The use of automated event annotation tools further opens new avenues for research that benefits from fast online annotation of events, such as during free recall. Real-time event annotation could be embedded in experiments and even combined with real-time neurofeedback, enabling more dynamic and interactive research approaches. Furthermore, because of variance in the subjective perception of events, studies need to collect a large number of participants to reliably annotate stimulus material, which makes annotated stimuli a scarce resource for experiments. Large language models make it possible to sample stimulus material from a large corpus of annotated narratives instead of relying on a few curated examples. Importantly, the variance between human annotators may contain meaningful information about individual human event perception; it is therefore important to note that, despite GPT-3's ability to capture human event perception, human behavioral data cannot be replaced as the gold standard for the investigation of human cognition. That is, studies that investigate individual segmentation behavior and are not interested in a consensus solution should refrain from using boundary annotations derived from Large Language Models. In a similar vein, systematic differences between human event annotations and annotations from large language models may become the subject of investigation.

Notably, the present results may represent a lower bound of the similarity between GPT-3-derived annotations and human event perception, because human participants in our study segmented the audio form of the narrative, whereas GPT-3 only had access to the text itself. In the audio form of the narrative, the speaker may convey information about event structure via pauses and intonation that are absent from the text and they may obscure events by rapidly connecting consecutive sentences. Furthermore, GPT-3 performed event segmentation based on a full text-segment; therein, the model could use information from the current and next event to place an event boundary. In contrast, human annotators

Table 2 Number of events derived from GPT-3 output and the corresponding Hamming distance to the consensus human annotations

Story	N GPT events	distance / p value	N GPT *long* events	distance / p value
"Pieman (run 1)"	23	0.255 ($p = 0.024$)	14	0.223 ($p = 0.069$)
"Pieman (run 2)"		0.245 ($p = 0.009$)		0.191 ($p = 0.004$)
"Monkey"	88	0.25 ($p = 0.044$)	59	0.193 ($p = 0.024$)
"Tunnel"	139	0.264 ($p<0.001$)	76	0.168 ($p = 0.002$)

The event boundaries segmented by GPT-3 significantly align with the consensus human annotations for almost all tested settings, as indicated by p values < 0.05

had to make button-press decisions only based on past information (with the notable exemption of the second behavioral run of the "Pieman" story, where human participants could use their memory of the text to improve their segmentation). Conversely, GPT-derived boundaries could only be resolved at a sentence-level resolution. The attribution of human event boundaries to the next sentence boundary effectively treats the occurrence of mid-sentence boundaries as "noise" to be removed by the analysis, whereas some of these responses may be cognitive "signal" that is of interest to the experimenter. This is an important limitation of this approach. Future research may investigate the importance of such differences between GPT-3 and human perception further. Additionally, our evaluation uses datasets that are representative of those commonly studied in human cognitive science and has, for instance, been successfully applied in recent work to analyze a new text domain and dataset: free-form autobiographical memories from over 1000 participants (Rouhani et al., 2023). It is, however, important to note that those datasets are not representative of the entire world population (e.g., they are in English and are Western-focused). Researchers who want to build on our methods should consider the similarity of their experimental stimulus and participant population.

In this study, we tested prompt variations that changed the segmentation behavior of GPT-3 from a fine-grained segmentation to a coarser segmentation. This coarse segmentation appeared to match human behavior better. However, future work may explore if responses from human subjects that are instructed to segment *the smallest units* that seem natural to them (Zacks et al., 2007) align more closely with the GPT-derived segmentation without the *long* qualifier. It is also noteworthy that the *long* GPT-derived event boundaries are not always a subset of the small event boundaries, which can happen in GPT and humans, for instance, when a coherent train of thought spans two separate larger events (e.g., two time periods).

We also cannot rule out that – with further prompt engineering – a different prompt will provide an even more human-like segmentation than the prompt that we used. In the Supplementary material, we discuss some anecdotal observations on prompt variations. Notably, expanding the definition of event often resulted in text output that diverged from a word-for-word copy; removing the definition of "event", on the other hand, resulted in a segmentation that appeared to be based on the colloquial understanding of event as "occurrence" (e.g., a person stopping).

The demonstration that GPT-3 is able to segment a continuous narrative into events similarly to humans has implications for both cognitive science and machine learning. In cognitive science, researchers have already started viewing language models as useful *model organisms* for human language processing (Toneva, 2021) since they present one implementation of a language system that may be very different from that of the human brain, but may nonetheless offer insights into the nature of linguistic tasks and the computational processes that are sufficient (or insufficient) to solve them (McCloskey, 1991; Baroni, 2020). Our findings suggest that a large language model can be viewed not only as a model organism for language processing but also more generally for event cognition. On the other hand, even though GPT-3-derived event segmentation significantly aligns with average human behavior, there is still a substantial gap between the behavior of the model and that of the most consistent humans. Biasing the language model to internally represent even more human-like event boundaries may result in improved performance on machine learning tasks. For instance, prior work has already demonstrated that "chunking" continuous input can help a modified transformer model to improve its memory performance (Lampinen et al., 2021). Chunking narrative texts into events in a way that better mirrors human event perception may be a productive avenue to optimize cognitive tasks in machines.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-024-02569-z.

Acknowledgements The authors wish to thank Shailee Jain, Vy Ai Vo, Javier Turek, and Alex Huth for helpful comments and discussions.

Funding This work was funded by ONR MURI grant N00014-17-1-2961 awarded to KAN. MT was funded by the Princeton Neuroscience Institute CV Starr fellowship and NIMH T32 MH065214, and SM was funded by Deutsche Forschungsgemeinschaft (DFG), project 437219953.

Availability of data and materials All data that support this manuscript are publicly available from Zenodo (https://doi.org/10.5281/zenodo.10055827).

Code availability All code that produces the findings in this manuscript is publicly available from Zenodo (https://doi.org/10.5281/zenodo.10055827).

Declarations

Ethics approval All studies reported in this manuscript were reviewed and approved by the Princeton University Institutional Review Board.

Consent to participate Informed consent was obtained from all participants prior to participation in the studies in accordance with the requirements of the Princeton University Institutional Review Board.

Consent for publication The authors affirm that human research participants provided informed consent for publication of the anonymized data in accordance with the requirements of the Princeton University Institutional Review Board.

Competing interests The authors declare no competing interests.

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.e5. https://doi.org/10.1016/j.neuron.2017.06.041

Bangert, A. S., Kurby, C. A., Hughes, A. S., & Carrasco, O. (2020). Crossing event boundaries changes prospective perceptions of temporal length and proximity. *Attention, Perception, & Psychophysics, 82*(3), 1459–1472. https://doi.org/10.3758/s13414-019-01829-x

Baroni, M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B, 375*(1791), 20190307.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877–1901.

Brunec, I. K., Moscovitch, M., & Barense, M. D. (2018). Boundaries shape cognitive representations of spaces and events. *Trends in Cognitive Sciences, 22*(7), 637–650. https://doi.org/10.1016/j.tics.2018.03.013

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science., 6*, 3–5. https://doi.org/10.1177/1745691610393980

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). Gpt-2's activations predict the degree of semantic comprehension in the human brain. *bioRxiv*.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology, 5*(1), 1–10.

Chavis, D. (2017). So a monkey and a horse walk into a bar [Audio Podcast] [This American Life. https://www.thisamericanlife.org/631/so-a-monkey-and-a-horse-walk-intoa-bar/act-one-0].

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. https://doi.org/10.1038/nn.4450

Clewett, D., DuBrow, S., & Davachi, L. (2019). Transcending time in the brain: How event memories are constructed from experience. *Hippocampus, 29*(3), 162–183.

DuBrow, S., & Davachi, L. (2013). The influence of context boundaries on memory for the sequential order of events. *Journal of Experimental Psychology. General, 142*(4), 1277–1286. https://doi.org/10.1037/a0034024

DuBrow, S., & Davachi, L. (2016). Temporal binding within and across events. *Neurobiology of Learning and Memory, 134*, 107–114. https://doi.org/10.1016/j.nlm.2016.07.011

Ezzyat, Y., & Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological Science, 22*(2), 243–252. https://doi.org/10.1177/0956797610393742

Golan, T., Siegelman, M., Kriegeskorte, N., & Baldassano, C. (2022). Testing the limits of natural language models for predicting human language judgments. arXiv:2204.03592.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience, 25*(3), 369–380.

Hamming, R. W. (1950). Error detecting and error correcting codes [Conference Name: The Bell System Technical Journal]. *The Bell System Technical Journal, 29*(2), 147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x

Jafarpour, A., Buffalo, E. A., Knight, R. T., & Collins, A. G. (2022). Event segmentation reveals working memory forgetting rate. iScience, 25 (3), 103902. https://doi.org/10.1016/j.isci.2022.103902

Jain, S., & Huth, A. (2018). Incorporating context into language encoding models for fMRI. *Advances in Neural Information Processing Systems, 31*, 6628–6637.

Kumar, M., Goldstein, A., Michelmann, S., Zacks, J. M., Hasson, U., & Norman, K. A. (2022). Bayesian Surprise Predicts Human Event Segmentation in Story Listening. https://doi.org/10.31234/osf.io/qd2ra

Kurby, C. A., & Zacks, J. M. (2011). Age differences in the perception of hierarchical structure in events [Place: Germany Publisher: Springer]. *Memory & Cognition, 39*(1), 75–91. https://doi.org/10.3758/s13421-010-0027-2

Lampinen, A. K., Chan, S., Banino, A., & Hill, F. (2021). Towards mental time travel: A hierarchical memory for reinforcement learning agents. *Advances in Neural Information Processing Systems, 34*, 28182–28195.

Lositsky, O., Chen, J., Toker, D., Honey, C. J., Shvartsman, M., Poppenk, J. L., Hasson, U., & Norman, K. A. (2016). Neural pattern change during encoding of a narrative predicts retrospective duration estimates (H. Eichenbaum, Ed.). *eLife*, *5*, e16070. https://doi.org/10.7554/eLife.16070

Lu, Q., Hasson, U., & Norman, K. A. (2022). A neural network model of when to retrieve and encode episodic memories. eLife, 11, e74445. https://doi.org/10.7554/eLife.74445

Mariola, A., Fountas, Z., Barnett, L., & Roseboom, W. (2022). Event segmentation in continuous, naturalistic videos from model-based, data-driven, and human perspectives. https://doi.org/10.31234/osf.io/t6fkg

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological science, 2*(6), 387–395.

Michelmann, S., Hasson, U., & Norman, K. A. (2023). Evidence That Event Boundaries Are Access Points for Memory Retrieval. *Psychological Science, 34*(3), 326–344. https://doi.org/10.1177/09567976221128206

Michelmann, S., Price, A. R., Aubrey, B., Strauss, C. K., Doyle, W. K., Friedman, D., Dugan, P. C., Devinsky, O., Devore, S., Flinker, A., Hasson, U., & Norman, K. A. (2021). Moment-by-moment tracking of naturalistic learning and its underlying hippocampo-cortical interactions. *Nature Communications, 12*(1), 5394. https://doi.org/10.1038/s41467-021-25376-y

O'Grady, J. (2008). Pieman [Audio Podcast] [The Moth. https://themoth.org/stories/pie-man].

Pettijohn, K. A., & Radvansky, G. A. (2016). Narrative event boundaries, reading times, and expectation. *Memory & Cognition, 44*(7), 1064–1075. https://doi.org/10.3758/s13421-016-0619-6

Pohl, F. (1956). The Tunnel under the World [Radio Broadcast] [X Minus One. https://www.oldtimeradiodownloads.com/sci-fi/x-minus-one/tunnel-under-the-world1956-03-14].

Roseboom, W., Seth, A., Sherman, M., & Fountas, Z. (2022).*The Perception of Time in Humans, Brains, and Machines*. https://doi.org/10.31234/osf.io/c7vzx

Rouhani, N., Stanley, D., Team, C.-D., Adolphs, R. (2023). Collective events and individual affect shape autobiographical memory. *Proceedings of the National Academy of Sciences, 120*(29), e2221919120.

Salvador, S., & Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal., 11*(5), 561–580.

Sargent, J. Q., Zacks, J. M., Hambrick, D. Z., Zacks, R. T., Kurby, C. A., Bailey, H. R., Eisenberg, M. L., & Beck, T. M. (2013). Event segmentation ability uniquely predicts event memory. *Cognition, 129*(2), 241–255. https://doi.org/10.1016/j.cognition.2013.07.002

Sasmita, K., & Swallow, K. M. (2023). Measuring event segmentation: An investigation into the stability of event boundary agreement across groups. *Behavior Research Methods, 55*(1), 428–447. https://doi.org/10.3758/s13428-022-01832-5

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences, 118*(45), e2105646118.

Shin, Y. S., & DuBrow, S. (2021). Structuring Memory Through Inference-Based Event Segmentation. *Topics in Cognitive Science, 13*(1), 106–127. https://doi.org/10.1111/tops.12505

Silva, M., Baldassano, C., & Fuentemilla, L. (2019). Rapid memory reactivation at movie event boundaries promotes episodic encoding. *The Journal of Neuroscience, 39*(43), 8538–8548. https://doi.org/10.1523/JNEUROSCI.0360-19.2019

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in Cognitive Sciences, 23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Sun, C., Yang, W., Martin, J., & Tonegawa, S. (2020). Hippocampal neurons represent events as transferable units of experience. *Nature Neuroscience, 23*(5), 651–663. https://doi.org/10.1038/s41593-020-0614-x

Toneva, M. (2021). Bridging language in machines with language in the brain [Doctoral dissertation, Carnegie Mellon University].

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, *32*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need [arXiv:1706.03762 [cs]]. https://doi.org/10.48550/arXiv.1706.03762. Comment: 15 pages, 5 figures.

Wang, Z., Jafarpour, A., & Sap, M. (2022). Uncovering surprising event boundaries in narratives. *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, 1–12. https://doi.org/10.18653/v1/2022.wnu-1.1

Yuan, J., & Liberman, M. (2008). Speaker identification on the scotus corpus. *The Journal of the Acoustical Society of America, 123*, 3878–3878. https://doi.org/10.1121/1.2935783

Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in reading and film comprehension. *Journal of experimental psychology. General, 138*(2), 307–327. https://doi.org/10.1037/a0015305

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin, 133*(2), 273–293. https://doi.org/10.1037/0033-2909.133.2.273

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.