The Role of Episodic Memory in Storytelling: Comparing Large Language Models with Humans

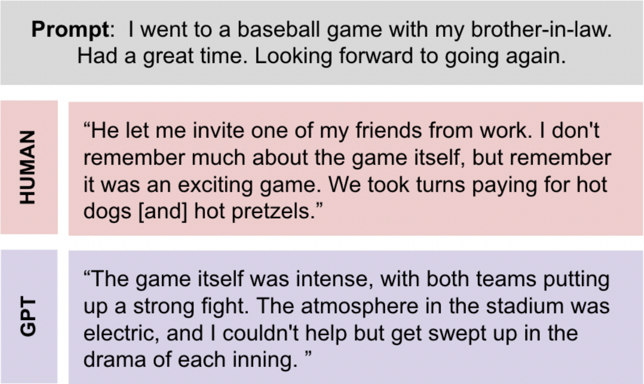
Charlotte A. Cornell [a,*], Shuning Jin [b,*], Qiong Zhang [a,b]
{charlotte.cornell, shuning.jin, qiong.z}@rutgers.edu
[a] Department of Psychology, [b] Department of Computer Science,
Rutgers University–New Brunswick, Piscataway, NJ, USA 08854

Abstract

We compare storytelling in GPT-3.5, a recent large language model, with human storytelling. Although GPT models are capable of solving novel and challenging tasks and matching human-level performance, it is not well understood if GPT processes information similarly as humans. We hypothesized that GPT differs from humans in the kind of memories it possesses, and thus could perform differently on tasks influenced by memory, such as storytelling. Storytelling is an important task for comparison as GPT becomes an increasingly popular writing and narrative tool. We used an existing dataset of human stories, either recalled or imagined (Sap et al., 2022), and generated GPT stories with prompts designed to align with human instructions. We found that GPT's stories followed a common narrative flow of the story prompt (analogous to semantic memory in humans) more than details occurring in the specific context of the event (analogous to episodic memory in humans). Furthermore, despite lacking episodic details, GPT-generated stories exhibited language with greater word affect (valence, arousal, and dominance). When provided with examples of human stories (through few-shot prompting), GPT was able to align its stories' narrative flow with human imagined stories but not human recalled stories. GPT was unable to match its affective aspects with either human imagined or recalled stories. We discuss these results in relation to GPT's training data as well as the way it was trained.

Keywords: episodic memory, storytelling, narrative flow, word affect, large language models

GPT (Generative Pre-trained Transformer) is a family of large language models trained to predict the next word on a vast amount of text corpora (T. Brown et al., 2020; Bubeck et al., 2023). This extensive training, along with the increased complexity of more recent model architectures, has allowed GPT models to perform far beyond mere text generation and conversational skills. For example, the latest GPT models can solve novel and challenging tasks that span mathematics, coding, law, psychology, and more (Bubeck et al., 2023; Singhal et al., 2023; Katz, Bommarito, Gao, & Arredondo, 2023). Despite its ability to produce human-like texts and achieve human-level performance, we do not know very well if GPT learns, thinks, and decides like humans (Binz & Schulz, 2023). In the current work, we hypothesize that GPT differs from humans in an important aspect of cognition – episodic memory, which could produce differences in the way GPT tells stories compared with humans.



Prompt: I went to a baseball game with my brother-in-law. Had a great time. Looking forward to going again.

HUMAN

"He let me invite one of my friends from work. I don't remember much about the game itself, but remember it was an exciting game. We took turns paying for hot dogs [and] hot pretzels."

GPT

"The game itself was intense, with both teams putting up a strong fight. The atmosphere in the stadium was electric, and I couldn't help but get swept up in the drama of each inning. "

Figure 1: Excerpts from a human recalled story and a GPT-generated story. Human stories include autobiographical details drawn from episodic memory in addition to semantic details mentioned by both GPT and humans. We hypothesize that GPT possesses primarily semantic memory, and therefore, writes more generic stories.

In humans, information retrieved from the past can be divided into two distinct types, episodic memory and semantic memory (Tulving, 1972). Episodic memory refers to our ability to remember individual experiences or events that occurred in particular spatial and temporal contexts; in contrast, semantic memory refers to general knowledge we have about the world, formed across multiple experiences (Squire, 1992; Tulving, 1972; Squire & Zola, 1998). Despite GPT's training data containing a wide variety of events and knowledge that could provide the basis for episodic memory, we hypothesize that GPT possesses primarily semantic memory, as it is trained to extract regularities from the training data rather than memorize from it (McCoy, Smolensky, Linzen, Gao, & Celikyilmaz, 2023). This is consistent with past observations where GPT is capable of predicting factual statements about the world, such as predicting "Seattle" given the prefix "The Space Needle is located in the city of" (Meng, Bau, Andonian, & Belinkov, 2022), but makes errors on factual information that it can only learn over one or a few occurrences (as much as 55% on bibliographic citations; Walters & Wilder, 2023). Thus it stands to reason that on tasks that rely strongly on episodic details, like storytelling, GPT would behave dif-

[*] CAC and SJ contributed equally and should be considered co-first authors

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

ferently. For example, when GPT generates a story about a baseball game, we predict its narrative to follow a typical sequence of events, such as the intensity of the game leading to an "electric" atmosphere and getting "swept up in the drama" (see Figure 1). When humans recall an event though, we rely both on common knowledge, such as the game being exciting (Gilboa, Rosenbaum, & Mendelsohn, 2018; Graesser, Robertson, & Anderson, 1981; Hyman Jr & Loftus, 1998), as well as autobiographical details about the event such as "invite one of my friends from work", "paying for each other's hot dogs and hot pretzels", and "not remembering much about the game itself" (Conway, Collins, Gathercole, & Anderson, 1996; Tulving, 1972).

As GPT becomes increasingly popular as a writing tool, it is important to understand how it writes stories compared to human storytelling. Recent research has shown that GPT can write coherent stories (Fan, Lewis, & Dauphin, 2018; See, Pappu, Saxena, Yerukola, & Manning, 2019) and collaboratively add to stories, back and forth with human respondents (Branch, Mirowski, & Mathewson, 2021; Nichols, Gao, & Gomez, 2020). As episodic memory heavily influences how humans tell stories (Conway et al., 1996; Conway, Pleydell-Pearce, Whitecross, & Sharpe, 2003), we hypothesize that there are important differences between GPT and humans in storytelling related to the role of episodic memory. One way in which psychological studies have previously analyzed the engagement of episodic memory is narrative flow. Sap et al. (2022) recently developed a metric of *sequentiality* to compare human recall and imagination, reasoning that human imagination relies more on semantic knowledge where each sentence depends strongly on the prior sentences. Autobiographical stories, however, have a less predictable sequence as episodic memories can deviate from the semantic knowledge associated with an event (Zwaan & Radvansky, 1998; Reichardt, Polner, & Simor, 2020). Another way episodic memory has been analyzed is to measure its affective content, as autobiographical memories with episodic details are associated with higher affect (Berntsen & Rubin, 2002; R. Brown & Kulik, 1977). We propose to capture the affective content of a story's language by using three of the most important, and largely independent, dimensions of word meaning: *valence* (positiveness–negativeness), *arousal* (active–passive), and *dominance* (dominant–submissive). Large VAD lexicons made possible through crowdsourcing allow us to obtain word ratings for words in prose-like text (e.g., Mohammad, 2018). As we hypothesize that GPT lacks the kind of episodic details observed in human stories, we predict that, GPT (1) will write stories more sequentially and (2) will use language that is low in word affect compared with human storytelling.

To test our hypotheses, we follow human experiments by prompting GPT to tell stories similarly to how humans were prompted in a storytelling task (Sap et al., 2022). We further provide GPT with examples of how humans wrote stories, both when recalling events and imagining them, to see if GPT can write stories more similarly to humans (i.e., few-shot prompting). To foreshadow our results, we find that GPT-generated stories are more sequential than human stories (especially human recalled stories), supporting our hypothesis that GPT-generated stories contain fewer episodic details than humans. Contrary to our expectations, we also find that GPT-generated stories have higher VAD ratings along all three word affect dimensions. Upon providing GPT with examples of human stories (through few-shot prompting), we find that GPT can align its stories' narrative flow with human imagined stories but not human recalled stories. GPT cannot match its affective aspects with either human imagined or recalled stories.

Method

In this section, we first overview Sap et al. (2022)'s HIPPOCORPUS dataset which contains stories that humans recalled and imagined. We then discuss how we created datasets of GPT-generated stories. Finally, we discuss the two measures used to analyze these datasets: sequentiality (Sap et al., 2022) and word affect (Mohammad, 2018).

HIPPOCORPUS Dataset

The HIPPOCORPUS dataset contains 6,854 stories that humans recalled, imagined, or retold (Sap et al., 2022). In the recalled condition, participants wrote a 15-25 sentence story about a memorable event they experienced in the past 6 months and then summarized it in 2-3 sentences. In the imagined condition, a second group of participants received one of those summaries and wrote a 15-25 sentence diary-like entry imagining that event. We did not use the retold stories. We only included stories for which there was both a recalled event and an imagined event in the dataset, and we excluded any repeated imagined stories (as sometimes different participants in the imagined condition were assigned the same recalled prompt). This left us with $N = 2,572$ recalled stories and $N = 2,572$ imagined stories.

GPT Datasets

Zero-shot Prompting Within the GPT family, we focused on GPT-3.5, based on InstructGPT (Ouyang et al., 2022) that extends GPT-3 (T. Brown et al., 2020). We prompted GPT with human summaries from recall stories (similarly to how human participants generated imagined stories based on summaries), giving us $N = 2,572$ GPT-generated stories. While humans saw an entire page of instructional text, the instructions we gave to GPT were a condensed version, focusing on the key instructions humans were given and controlling for the length of output: *"Given a short prompt summary, write an imagined story about an event. Write using a first-person perspective. The story must have at least 15 sentences and at most 25 sentences. The story must have at least 120 words and at most 600 words."* GPT saw the full prompt template as follows: *"{instructions} Summary: {summary} Story:"*, and generated the corresponding story as continuation. For implementation details, we used the "gpt-3.5-turbo" endpoint through OpenAI's API and used greedy decoding (i.e., generate the next most probable token).

Few-shot Prompting Beyond testing how GPT tells stories without any learning, we also wanted to explore if GPT would align its stories to be more like human stories if it was provided with human story examples. Few-shot prompting (i.e., providing a sequence of input-output pairs as the task demonstration) can enable GPT to generate higher-quality answers by prompting it with example responses (T. Brown et al., 2020). The effectiveness is rooted in a surprising ability of GPT: it can learn from the examples on the fly without explicit training (i.e., fine-tuning of model parameters). To do this, we randomly sampled 40 example stories from the imagined dataset (as the API allowed for a maximum of 16,385 tokens per prompt). We then prompted GPT with the same instructions accompanied by these 40 examples within the prompt. Specifically, the few-shot prompt template was: *"{instructions} {examples} {instructions} Summary: {summary} Story:"*. We repeated the instructions twice because the model tended to deviate from the instructions after seeing the long sequence of examples. For each story generation, we used the same samples but randomized the order (as language models use information from different positions of a long prompt differently; Liu et al., 2023). In addition to human examples from the imagined dataset, we separately repeated the above process by prompting GPT with human examples from the recalled dataset.

Linguistic Measures

Sequentiality We quantified a response's narrative flow by following Sap et al. (2022)'s metric, *sequentiality*. Given a topic summary and a story, it considers how each sentence in the story is causally determined by the topic and its preceding sentences. Intuitively, a sentence of high sequentiality is mainly driven by its preceding sentences, whereas one with low sequentiality is driven by the topic itself. Formally, the sequentiality of the i-th sentence (s_i) is the difference between two log-likelihoods, normalized by the sentence length:

$$\textsc{Seq}(s_i) = \frac{1}{|s_i|}[\log \underbrace{p_{\text{LM}}(s_i \mid \mathcal{T}, s_{<i})}_{\text{contextual}} - \log \underbrace{p_{\text{LM}}(s_i \mid \mathcal{T})}_{\text{topical}}].$$

The first term is log-likelihood under a contextual model (conditioned on the summary $\mathcal{T}$ and all prior sentences $s_{<i}$), and the second term is the log-likelihood under a topical model (conditioned on the summary $\mathcal{T}$ alone). Following Sap et al. (2022), we used a neural language model to measure these likelihoods, i.e., GPT-3. The sequentiality of a story is the average score of its sentences.

Word Affect We measured the affective content of stories using the NRC-VAD lexicon which contains *valence*, *arousal*, and *dominance* ratings for over 20,000 English words (Mohammad, 2018). VAD ratings range from 0 to 1 and higher ratings correspond to language that is more positive, more active, and more dominant (respectively). For each story, we calculated the average word rating for each dimension. Stop words, such as "and" or "the" or "my" (Nothman, Qin, & Yurchak, 2018), do not contain much affective information and were removed from the analysis (56% of human stories and 54% of GPT stories) as were any other words not covered in the NRC-VAD lexicon (8% of words in human stories and 6% in GPT stories).

Results

In this section, we compare human recalled and imagined stories with GPT-generated stories. We first examine GPT's stories that responded to a prompt similar to what human imagined group saw (zero-shot prompting). Then, we compare GPT's results with alternative prompts to test the robustness of the results. Last, we examine GPT's stories when GPT was given both our original instructions along with 40 examples of human stories (few-shot prompting).

Zero-shot Prompting

Figure 2A displays the sequentiality scores in the human recalled, human imagined, and GPT-generated stories. We first replicated Sap et al. (2022)'s results for human stories: Using a paired-samples t-test, where recalled and imagined stories were paired if sharing the summary prompt, we observed that imagined stories were significantly more sequential than recalled stories ($t(2571) = 31.93, p < .001$), suggesting that recalled stories contain more episodic details and flow in a less expected manner than imagined stories. When prompting GPT with instructions similar to human participants', its stories had greater sequentiality than human recall ($t(2571) = 52.84, p < .001$), supporting our hypothesis that GPT relies primarily on semantic memory and writes stories with greater narrative flow, as opposed to human recall which draws on episodic details. We do not have a strong hypothesis on the comparison between GPT-generated stories and human imagined stories; nevertheless, we observed a small but significant difference between them ($t(2571) = 10.50, p < .001$), although this effect no longer holds when using alternative large language models (GPT-2 or Llama2-7B; Touvron et al., 2023) for evaluating sequentiality. Meanwhile, human recalled stories are robustly less sequential than GPT-generated stories and human imagined stories when using alternative large language models for evaluating sequentiality.

Regarding the average word affect within stories (Figure 2B–D), human imagined stories contained language that was more positive ($t(2571) = 2.21, p = .027$) than human recall, but there was no significant difference in measures of arousal ($t(2571) = 1.00, p = .32$) or dominance ($t(2571) = 0.64, p = .52$). In contrast to human stories, GPT-generated stories were more positive (recalled: $t(2571) = 28.97, p < .001$; imagined: $t(2571) = 23.82, p < .001$). Additionally, GPT had higher ratings than human stories along the other dimensions: its language was more active (recalled: $t(2571) = 19.30, p < .001$; imagined: $t(2571) = 17.17, p < .001$) and more dominant (recalled: $t(2571) = 29.14, p < .001$; imagined: $t(2571) = 26.08, p < .001$). This result was surprising

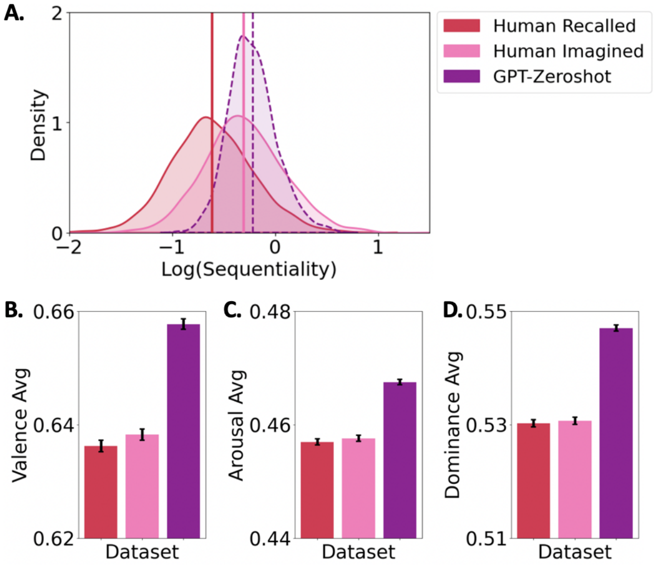


Figure 2: GPT-generated stories (A) were more sequential and (B) contained language that was more positive, (C) active, and (D) dominant than both human recall and imagination. Error bars represent the standard error of the mean (SEM).

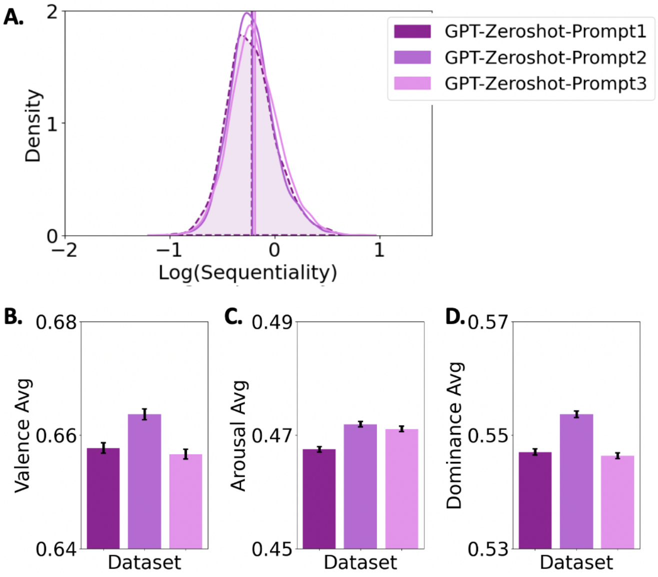


Figure 3: Our major conclusions about GPT's sequentiality and word affect were not overly sensitive to our original prompt's wording. (A) GPT's stories were more sequential with the new prompt alternatives than the original Prompt-1. (B) The average valence of GPT's language increased with Prompt-2 and did not change with Prompt-3. (C) Arousal and (D) dominance increased with both prompt alternatives. Error bars represent SEM.

as we predicted GPT to write with less affective language, and we later discuss this in the general discussion.

Next, we explored if GPT's results of higher sequentiality and VAD ratings were sensitive to certain words in the original instructional prompt it received (Prompt 1). For one, we explored if GPT was sensitive to the words "imagined story" as it may rely on exaggerated, fictional stories in its training data to inform its storytelling. Therefore, we changed these words to "event" (Prompt 2). Nonetheless, GPT-generated stories had significantly more sequentiality than with the original prompt ($t(2571) = 2.88, p = 0.004$) and had language that was more positive ($t(2571) = 11.98, p < .001$), more active ($t(2571) = 11.87, p < .001$), and more dominant ($t(2571) = 16.48, p < .001$) compared to the original prompt. Thus, using the word "event" instead of "story" did not align GPT's storytelling closer to human writing (see Figure 3). We also tried another prompt version (Prompt 3) in which we added the instruction to imagine a story that was "memorable or surprising", which the human recall group was also asked, but we did not include in the original prompt to reduce biases. However, again, the sequentiality of GPT-generated stories was significantly greater than the original prompt ($t(2571) = 10.55, p < .001$), the arousal rating of GPT's language increased ($t(2571) = 9.88, p < .001$), and the dominance rating did not change significantly ($t(2571) = -1.67, p = 0.095$). Interestingly, GPT used less positive language with this additional instruction ($t(2571) = -2.22, p = .027$), though it was still more positive than both human imagination ($t(2571) = 22.62, p < .001$) and human recall ($t(2571) = 27.51, p < .001$). In sum, our major conclusions about the narrative flow and word affect in GPT storytelling stay unchanged under alternative prompt wordings.

Few-shot Prompting

With our original and alternative prompts, GPT-generated stories consistently showed higher sequentiality and VAD ratings than humans. Next, we examined if additionally providing examples of human stories would help GPT align the language of its stories (both sequentiality and average word affect) closer to that of human stories. When providing 40 example imagined stories along with our original prompt, the sequentiality of GPT-generated stories significantly reduced from zero-shot prompting ($t(2571) = -18.84, p < .001$) such that it was similar to human imagination ($t(2571) = 1.94, p = .053$). When providing 40 example recalled stories, the sequentiality of GPT-generated stories was also lower than with zero-shot prompting ($t(2571) = -24.70, p < .001$); however, its sequentiality was still greater than observed in human recall ($t(2571) = 40.45, p < .001$). Thus, these story examples were effective at helping GPT write with less sequentiality to the extent of human imagination, but not effective enough for the narrative flow of GPT-generated stories to align with that of human recall (Figure 4A).

Further, when provided with human examples, the average word affect of GPT-generated stories increased (instead

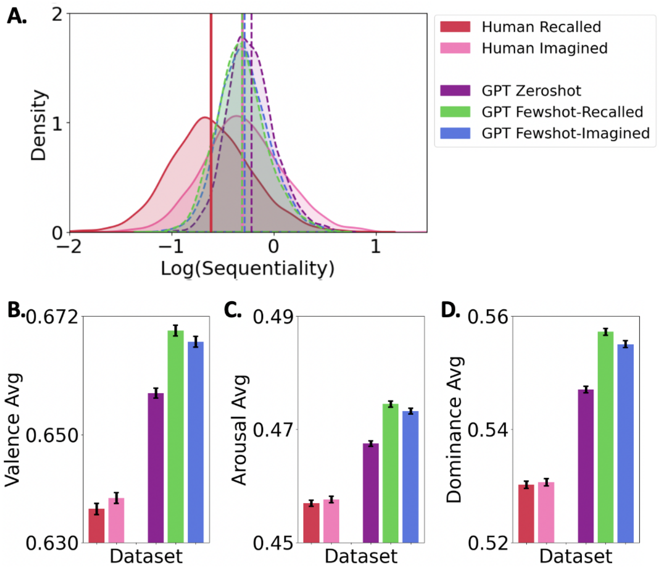


Figure 4: Effects of providing GPT with randomly-selected human story examples. (A) GPT wrote with less narrative flow when given examples than with zero-shot prompting. With imagined examples, GPT's sequentiality was similar to that of human imagination. GPT's language was more (B) positive, (C) active, and (D) dominant with few-shot prompting than with zero-shot prompting. Error bars represent SEM.

of decreased) with few-shot prompting compared to the zero-shot stories (Figures 4B–D). That is, after seeing 40 examples of imagined stories, GPT wrote with greater valence ($t(2571) = 18.64, p < .001$), arousal ($t(2571) = 15.04, p < .001$), and dominance on average ($t(2571) = 19.85, p < .001$); and when GPT received examples of recalled stories, the same trend held along valence ($t(2571) = 22.09, p < .001$), arousal ($t(2571) = 17.57, p < .001$) and dominance ($t(2571) = 24.22, p < .001$). This result is surprising because the examples of human stories had lower VAD ratings.

The few-shot prompting we used so far provided GPT with 40 randomly selected examples of imagined and recalled stories. Next, we explored if specifically selecting the 40 most (or least) sequential stories or the stories with the most (or least) valence would lead GPT to align its storytelling more closely to these more extreme examples (Figures 5A and 5C display the distributions of the sequentiality of the three sets of example stories in few-shot prompting, showing that the random stories were indeed centered between the most and least sequential stories). Compared to the 40 random example stories, the sequentiality of GPT-generated stories increased when providing the 40 most sequential stories (recalled: $t(2571) = 13.23, p < .001$; imagined: $t(2571) = 13.96, p < .001$) and decreased when providing the 40 least sequential stories (recalled: $t(2571) = -11.54, p < .001$; imagined: $t(2571) = -14.16, p < .001$), indicating that GPT was sensitive to the amount of narrative flow in the examples it saw (Figures 5B and 5D). While GPT's sequentiality was still greater than human imagination when prompted with the most sequential examples ($t(2571) = 7.99, p < .001$), the sequentiality of GPT-generated stories was less than that of human imagined stories when given examples of the least sequential stories ($t(2571) = -4.38, p < .001$). However, GPT was unable to write with less sequentiality to the extent of human recalled stories when we provided the most ($t(2571) = 46.10, p < .001$) and even the least ($t(2571) = 34.58, p < .001$) sequential examples of human recalled stories in the prompt. Taken together, while GPT did not write less sequentially to the extent of human recall, it was able to align its narrative flow with human imagination.

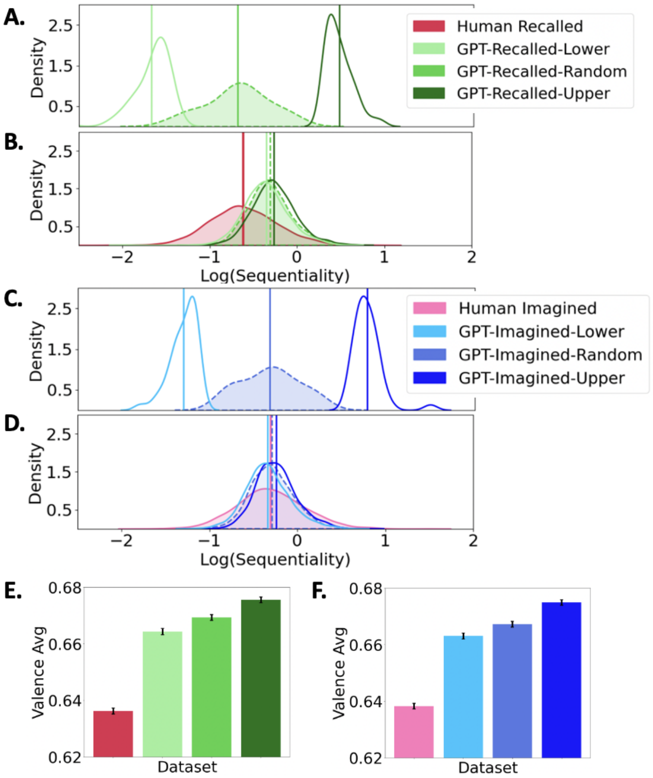


Figure 5: Effects of providing GPT with the most or least sequential or valenced human story examples. Distribution of the 40 random or most/least sequential examples from (A) recalled and (C) imagined stories. The sequentiality of GPT stories increased/decreased when provided with the most/least sequential stories; (B) it was still greater than human recalled stories but (C) similar to that of human imagined stories. The valence of GPT stories increased/decreased when provided with the most/least valenced (E) recalled or (F) imagined stories as examples, though it was still greater than observed in human stories. Error bars represent SEM.

We repeated a similar analysis by providing GPT with examples of stories with the highest or lowest average valence rating. Compared to few-shot prompting with 40 random examples of recalled stories, the average valence of GPT-generated stories increased when given the 40 stories with the greatest average valence ($t(2571) = 13.03, p < .001$) and decreased when given the 40 stories with the lowest average valence ($t(2571) = -10.18, p < .001$). Nonetheless, GPT-generated stories' average valence was still greater than in human recall stories when provided with the most ($t(2571) = 51.30, p < .001$) and least ($t(2571) = 36.73, p < .001$) valenced examples during few-shot prompting (Figure 5E). Similar results held when we compared the three sets of few-shot prompting with 40 imagined examples (Figure 5F). That is, compared to the 40 random examples of imagined stories, GPT's average valence increased when provided with the 40 most valenced stories ($t(2571) = 15.75, p < .001$) and decreased when provided with the 40 least valenced stories ($t(2571) = -8.55, p < .001$). Again, GPT's valence was still greater than human imagination with the most ($t(2571) = 43.16, p < .001$) and least ($t(2571) = 28.97, p < .001$) valenced examples. Taken together, GPT was sensitive to the average valence of the example stories; however, this selective few-shot prompting was not effective for GPT to write with affect to the same extent as human stories.

General Discussion

In this paper, we explored differences in storytelling between humans and GPT which may stem from episodic memory differences. We followed prior ways that episodic memory has been analyzed in human stories, namely narrative flow and word affect, and we compared GPT-generated stories to an existing dataset of human recalled and imagined stories. We found that without providing GPT with human story examples, GPT consistently wrote more sequentially and used language that, on average, had greater valence, arousal, and dominance ratings as compared to human stories (especially human recalled stories; see a summary of main results in Figure 6, visualizing this in the two-dimensional space of sequentiality scores and valence ratings). Providing GPT with human story examples assisted it in writing with similar sequentiality as human imagination but not human recall. However, human story examples were not effective at helping GPT to write with language lower in affect to the same extent as human stories. We now turn to a discussion of these results.

We hypothesized that, when storytelling, GPT is driven primarily by semantic memory as it has been trained to extract regularities from its training data (analogous to semantic memory in humans) rather than memorize from it (analogous to episodic memory in humans). Our finding that GPT generates stories with greater sequentiality than humans supports this hypothesis and suggests that GPT's stories followed an expected or common narrative flow of the story prompt (semantic memory) more than details occurring in the specific context of the event (episodic memory). Furthermore, as human imagined stories primarily rely on common knowledge and less so on episodic details, it is not surprising that, the difference in sequentiality between human imagined stories and GPT-generated stories is small, and that when GPT is provided with humans examples (through few-shot prompting), it was able to align its narrative flow with the human imagined stories but not human recalled stories.

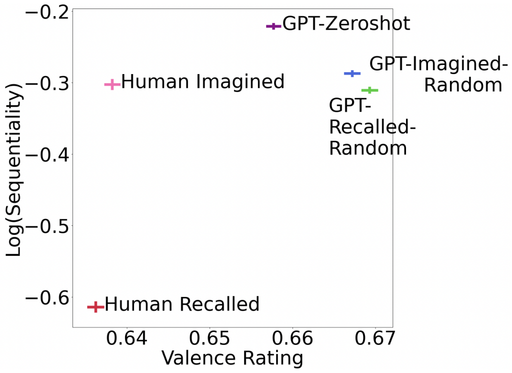


Figure 6: A summary of our main results. Log-transformed sequentiality scores and valence ratings for human recalled/imagined stories, as well as GPT zero-shot/few-shot prompting with examples from recalled/imagined stories. Error bars represent SEM.

We also hypothesized that GPT writes stories with language lower in affect, as it lacks the kind of episodic details human autobiographical memories have. However, we observed the opposite trend that GPT wrote stories with higher affect than human stories. Consider again the baseball story excerpt from Figure 1: GPT's story has language that was higher in affect (e.g., "intense", "electric", "swept up in the drama") whereas the human recalled story mentioned once that "it was an exciting game" but did not detail the emotion of the game. Though our sequentiality results suggest GPT-generated stories lacked episodic details, it might be trained with data high in affect. For instance, GPT's training data contains WebText2 (Radford et al., 2019; Kaplan et al., 2020), with links to external webpages upvoted by Reddit users that may include stories with language high in VAD ratings. Additionally, GPT-3.5 goes through supervised fine-tuning on human demonstration and reinforcement learning with human preference as a reward (Ouyang et al., 2022).

Prior work shows that few-shot prompting enables general-purpose language models to better recognize the task in hand and has the potential to override its pre-training prior (Wei et al., 2023; Pan, Gao, Chen, & Chen, 2023). However, even upon few-shot prompting using human recalled stories as examples, GPT was unable to write with a similar narrative flow. This result suggests that the way GPT was trained (in extracting regularities from its training data) limits its ability to write with low narrative flow as seen in humans recall. Similarly, GPT also was unable to align its average word affect with that of human stories upon seeing examples. However, we acknowledge that such an effect in word affect could be only an indication of GPT's training data about event knowledge (but not in general), as the human experiments we compared GPT with are centered around descriptions of events.

References

Berntsen, D., & Rubin, D. C. (2002). Emotionally charged autobiographical memories across the life span: The recall of happy, sad, traumatic and involuntary memories. *Psychology and Aging*, *17*(4), 636–652.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Branch, B., Mirowski, P., & Mathewson, K. W. (2021). Collaborative storytelling with human actors and AI narrators. *arXiv preprint arXiv:2109.14728*.

Brown, R., & Kulik, J. (1977). Flashbulb memories. *Cognition*, *5*(1), 73–99.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901).

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... others (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Conway, M. A., Collins, A. F., Gathercole, S. E., & Anderson, S. J. (1996). Recollections of true and false autobiographical memories. *Journal of Experimental Psychology: General*, *125*(1), 69–95.

Conway, M. A., Pleydell-Pearce, C. W., Whitecross, S. E., & Sharpe, H. (2003). Neurophysiological correlates of memory for experienced and imagined events. *Neuropsychologia*, *41*(3), 334–340.

Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (Vol. 1, pp. 889–898).

Gilboa, A., Rosenbaum, R. S., & Mendelsohn, A. (2018). Autobiographical memory: From experiences to brain representations. *Neuropsychologia*, *110*, 1–6.

Graesser, A. C., Robertson, S. P., & Anderson, P. A. (1981). Incorporating inferences in narrative representations: A study of how and why. *Cognitive Psychology*, *13*(1), 1–26.

Hyman Jr, I. E., & Loftus, E. F. (1998). Errors in autobiographical memory. *Clinical Psychology Review*, *18*(8), 933–947.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). GPT-4 passes the bar exam. *Available at SSRN 4389233*.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*.

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. (2023). How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*, *11*, 652–670.

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in gpt. In *Advances in neural information processing systems* (Vol. 35, pp. 17359–17372).

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 174–184).

Nichols, E., Gao, L., & Gomez, R. (2020). Collaborative storytelling with large-scale neural language models. In *Proceedings of the 13th acm siggraph conference on motion, interaction and games* (pp. 1–10).

Nothman, J., Qin, H., & Yurchak, R. (2018, July). Stop word lists in free open-source software packages. In *Proceedings of workshop for NLP open source software* (pp. 7–12).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in neural information processing systems.*

Pan, J., Gao, T., Chen, H., & Chen, D. (2023). What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the association for computational linguistics* (pp. 8298–8319).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Reichardt, R., Polner, B., & Simor, P. (2020). Novelty manipulations, memory performance, and predictive coding: The role of unexpectedness. *Frontiers in Human Neuroscience*, *14*, 152.

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.

See, A., Pappu, A., Saxena, R., Yerukola, A., & Manning, C. D. (2019). Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd conference on computational natural language learning* (pp. 843–861).

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... Payne, P. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180.

Squire, L. R. (1992). *Encyclopedia of learning and memory*. Macmillan.

Squire, L. R., & Zola, S. M. (1998). Episodic memory, semantic memory, and amnesia. *Hippocampus*, *8*(3), 205–211.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tulving, E. (1972). Episodic and semantic memory. *Organization of memory*, *1*, 381-403.

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by chatgpt. *Scientific Reports*, *13*(1), 14045.

Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., . . . Ma, T. (2023). Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185.