Playing Story Creation Games With Large Language Models: Experiments With GPT-3.5*

Timothy S. Wang and Andrew S. Gordon

University of Southern California, Los Angeles, CA USA
wangtimo@usc.edu, gordon@ict.usc.edu

Abstract. We created a web application where human users can play a story creation game with OpenAI's GPT-3.5, based on the Tell Tale card game. Tell Tale requires players to generate a brand new and coherent story based on a set of initial story elements, making the game a useful structure for exploring how well GPT-3.5 performs in generating coherent and engaging narratives. We show that GPT-3.5 performs remarkably well in generating such a narrative based on a random set of initial story elements, and that GPT-3.5 is even able to incorporate other literary elements such as suspense and flashbacks into its stories to enhance them and make them more engaging. By having human testers play Tell Tale with GPT-3.5 through our web application, we also demonstrate GPT-3.5's strong potential to be used as an interactive storytelling system, one that can both write and evaluate different narratives. We evaluate this potential using both quantitative and qualitative data from the human testers. Results indicate that, while GPT-3.5's narrative abilities are far from perfect, large language models have great potential in many different automated narrative situations.

Keywords: Narrative Generation · Narrative Evaluation · Large Language Models

1 Introduction

While human storytelling is universally seen across all cultures, it is by no means a trivial process. In telling a compelling story, humans must interpret together a set of outside observations and a set of initial world knowledge into a coherent narrative structure of what happened (an interpretation of events), before communicating that structure to others in a meaningful and coherent manner. One key area of current artificial intelligence and natural language processing research is automating this storytelling process, a feat which could greatly aid in integrating computers into this critical usage of human language, amongst other

* The project or effort depicted was or is sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005, and that the content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

potential applications. While certain aspects of this process, such as generating coherent text via computational linguistics, have seen enormous progress, historically being able to synthesize observations and world knowledge together into a meaningful narrative structure has been more challenging. Despite this, the introduction of large language models (LLMs) such as OpenAI's Generative Pre-trained Transformer 3.5 (GPT-3.5) [8] has provided researchers with a new and extremely powerful tool to tackle this challenge.

This paper explores the capability of GPT-3.5 to do competently automated narrative interpretation and, ultimately, general automated storytelling. To do this, GPT-3.5 was tested within the framework of the Tell Tale card game[1], a casual card game where players must generate an original narrative based on a random selection of 120 different story elements printed on 60 double-sided cards. The number of story elements used in a game can vary, but players must utilize some aspect of each of their selected story elements in their narrative. For example, if a player's cards are a baseball player, a heart, and a train, as depicted in Figure 1, one potential narrative could be: "The girl took the train to see the baseball game because she loves baseball." This narrative is acceptable because it mentions baseball, trains, and heart's association with love, but the story may not be judged by other players as being creative, clever, or well-crafted.

This research tested three tasks for GPT-3.5 to try to complete. First is the ability of GPT-3.5 to generate a coherent story from a random selection of Tell Tale story elements. Second is the ability of GPT-3.5 to complete the first task while also being able to incorporate into its narrative literary devices such as suspense, foreshadowing, and imagery. Third is the ability of GPT-3.5 to evaluate and score its own and other narratives within the Tell Tale game. In testing these three tasks, we developed a new Tell Tale web application where users can play a game of Tell Tale cards with GPT-3.5. We used this web application to conduct human evaluations of GPT-3.5's abilities.

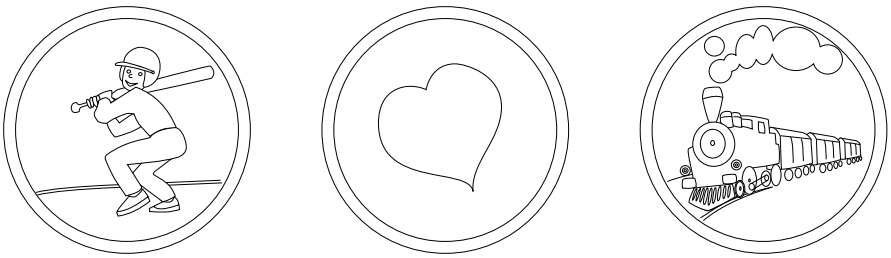

Fig. 1. Images of three Tell Tale cards, from [3]

[1] https://blueorangegames.com/index.php/games/telltale

In the following sections, we will show results that highlight GPT-3.5's impressive interpretive and storytelling abilities, with GPT-3.5 having success in both generating Tell Tale narratives and evaluating other Tell Tale narratives. Using simple prompting strategies, we were able to develop an interactive web application where users can "compete" against GPT-3.5 in a game of Tell Tale, revealing GPT-3.5's potential not just in generating coherent narratives, but also in interactive storytelling settings. We will also show preliminary quantitative and qualitative human test data on GPT-3.5's storytelling and story evaluation abilities, revealing initial human testers' generally positive evaluations of GPT-3.5's abilities.

2 Related Work

Gordon and Spierling [3] previously investigated automated narrative generation in story creation games, specifically looking at the Tell Tale card game. In their work, they generated eight variations of stories that incorporated elements from three Tell Tale cards (a baseball player, a train, and a heart symbol) using a logical abduction and a hand-crafted knowledge base of first-order logic axioms. The output of their system is a graph structure representing the story, much in the same vein as Elson's Story Intention Graphs [1] or the Causal Network models of Trabasso and van den Broek [11], which can be coupled with dedicated text-generation methods to produce fluent narratives [4]. Our work differs from theirs by utilizing a large language model for the entire narrative generation process, avoided the need for hand-authored knowledge bases or narrative planning algorithms.

The use of neural networks for narrative text generation has received enormous research attention over the years [7], as have methods for evaluating the output of various systems [9]. In more recent work, large language models are fine-tuned specifically for the narrative generation task. Representative of these newer models are MPT-7B-StoryWriter-65k+ [6], a decoder-style transformer fine-tuned on a fiction subset of the books3 dataset [2], and MythoMax-L2-13b[2], which targets storytelling and role-playing by mixing tensors from multiple upstream models based on Meta's LLaMa model [10]. Although these newer models may have characteristics that are well-suited for our narrative generation tasks, we investigate only OpenAI's popular GPT-3.5 model in our research.

Finally, there has also been recent work in enabling computer systems to incorporate certain narrative devices into their stories. For instance, Han et al. [5] used a Plan-and-Write framework and reinforcement learning to better incorporate the literary device of flashbacks into generated narratives, resulting in more interesting but still coherent stories. This paper also involves asking a computer system to incorporate flashbacks into its generated stories, but we use a pre-trained LLM (GPT-3.5) instead, and we also ask GPT-3.5 to incorporate a broad set of literary devices, including imagery and conflict, beyond just only flashbacks.

[2] https://huggingface.co/Gryphe/MythoMax-L2-13b

3 Prompt Engineering

Participation in a standard Tell Tale card game requires both the generation of a novel story and the consideration of stories created by other participants. To enable GPT-3.5 to execute both functions, we engineered specific prompts for three different tasks. Prompts were engineered by making individual API calls to GPT-3.5 Turbo, using a consistent temperature value of 0.7 for all API calls. In engineering these prompts, we focused our efforts on a small subset of Tell Tale cards (heart, baseball, train, star, and city). As our interest is on narrative text generation rather than visual understanding, these Tell Tale cards were presented to GPT-3.5 as textual labels, rather than as images that appear on the cards.

In the first task, we investigated whether GPT-3.5 could successfully perform the basic narrative generation task. As expected, even with simple prompts, GPT-3.5 had no difficulties in generating coherent narratives that successfully use the Tell Tale cards as story elements. For example, when prompted to write a story using "heart" and "star" as Tell Tale cards:

Prompt: *Write a coherent story that uses both some interpretation of "heart" and some interpretation of "star".*
Narrative (Abbreviated for space): *Once upon a time, in a small village nestled amidst a dense forest, there lived a young girl named Luna. Luna had a heart full of dreams and a passion for the stars that glittered in the night sky. She would spend countless hours staring at them, wondering about the mysteries they held.Luna's parents, who were simple farmers, noticed their daughter's fascination with the stars...*

In the second task, we explored whether GPT-3.5 could perform the narrative generation task while including certain literary features such as suspense, foreshadowing, and imagery. The motivation behind this second task is to move beyond simple interpretation and to explore whether GPT-3.5 can create narratives similar to how human writers tend to generate stories, with literary features that enhance the storytelling experience, making it more exciting and entertaining. Prompting GPT-3.5 to include literary devices such as foreshadowing proved to be more difficult. We quickly found that simply asking GPT-3.5 to include a certain literary device was not enough. Rather, GPT-3.5 was much more likely to include that literary device when we included a definition of that specific literary device. For instance:

Prompt: *Write a coherent story that uses both some interpretation of "city" and some interpretation of "baseball". The story must have narrative foreshadowing. The story having narrative foreshadowing means the story should have some object, situation, or plot point that hints at what will happen in the future."*
Narrative: (Abbreviated for space): *In the heart of the bustling city, where towering skyscrapers kissed the sky, there existed a hidden gem of a baseball field known as Diamond Park. Tucked away amidst the concrete jungle, it was an oasis for those seeking solace from the chaotic city life. Young Tommy, a passionate*

baseball enthusiast, spent his days at Diamond Park. He was a prodigious player, with a natural talent for the game that caught the attention of all who watched him. The city became captivated by Tommy's skills, and his reputation as a future baseball star grew with each game he played. One fateful day, as the sun set over the city, Tommy's favorite bat, lovingly named "Slugger," mysteriously disappeared from his locker. Heartbroken and devastated, he searched every nook and cranny of Diamond Park, but to no avail...

There were some literary devices that were more difficult to prompt GPT-3.5 to include in its narrative. For instance, GPT-3.5 had a particular difficulty in consistently including flashbacks into its stories when prompted. In addition, even if GPT-3.5 were able to technically include a literary device, sometimes the use of the literary device does not cleanly fit into the narrative and reads to be somewhat contrived. Overall though, these initial results generally showed GPT-3.5's ability to create coherent narratives based only on an initial set of Tell Tale cards.

In the third task, we looked at whether GPT-3.5 could also successfully evaluate other Tell Tale narratives, and score them based on how well they were able to incorporate all the Tell Tale cards that were used as observations. The motive behind this third task is to see if GPT-3.5 could emulate an actual human player in a game of Tell Tale, where players both have to come up with an original story based on a set of Tell Tale cards but also evaluate their story as well. This would allow us to both test GPT-3.5's ability to analyze a new story and GPT-3.5's potential as an interactive storytelling system. We found that reasonable results could be obtained by asking GPT-3.5 to evaluate how well its own generated stories adhered to the prompt, in the following format:

Prompt: *Evaluate the following story based on how well it follows the given prompt and give a score out of ten. The Prompt:* [Original Prompt]. *The Story:* [GPT-3.5's Story].

While not perfect, GPT-3.5's performance gave us the confidence that it could fully support an interactive Tell Tale card game web application.

4 Designing a Tell Tale Card Game Web Application

To test all three of these tasks, we created a simple web application where users can "compete" against GPT-3.5 in a game of Tell Tale cards, as shown in Figure 2. A round of this game involves a single human user interacting with the web interface to play against a single instance of GPT-3.5. The web application selects three random Tell Tale cards (from 120 possibilities) that serve as the starting "observations" for generating a narrative. In addition, the user is presented with seven literary features – conflict, suspense, foreshadowing, flashbacks, plot twists, personification, and imagery – and can choose any subset of those seven literary features. Based on the three Tell Tale cards provided to the user and the selected literary features, the user will then be asked to write a narrative involving those Tell Tale cards that also incorporates the selected literary features.

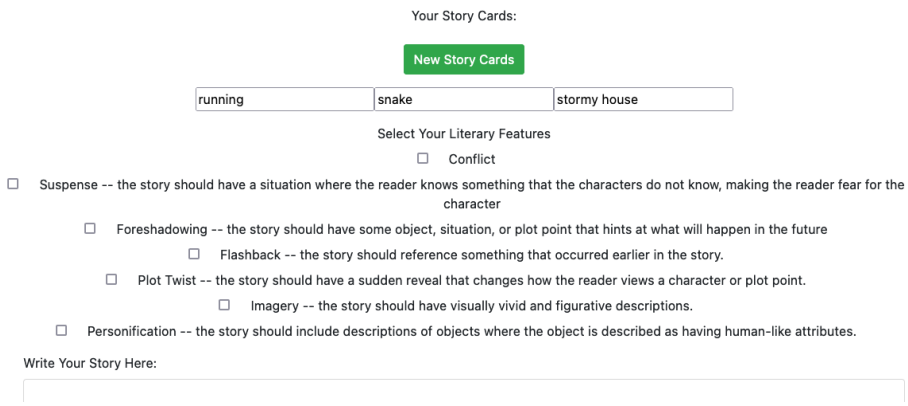

Fig. 2. Portion of the web application interface.

After the user submits their story, GPT-3.5 first writes a story itself using the same Tell Tale cards and selected literary features, using the following prompt template:

Prompt: *Write a coherent story that uses some interpretation of* [Tell Tale Card 1]*, some interpretation of* [Tell Tale Card 2]*, and some interpretation of* [Tell Tale Card 3]*. Also the story must have* [Literary Device 1]*. The story having* [Literary Device 1] *means the story should* [Definition of Literary Device 1]*. Also the story must have* [Literary Device 2]*...*

After generating a story, GPT-3.5 will then send two new and separate API requests, where GPT-3.5 will be prompted to evaluate and score both the user's story and its own story based on how well it incorporates the three Tell Tale cards and the selected literary devices:

Prompt: *Evaluate the following story based on how well it follows the given prompt and give a score out of ten. The Prompt:* [Original Prompt]*. The Story:* [User's or GPT-3.5's Story]*.*

By creating this web application, we hope to showcase the potential for GPT-3.5 as an interactive narrative agent, being able to both generate stories and to react intelligently to other narratives. In this way, this web application could be considered an initial step to using GPT-3.5 in a more complex interactive narrative system. Moreover, this web application also creates a natural way to test all three of our target tasks using human testers. By having human users interact with GPT-3.5 through our web application, the human users can provide evaluation data and feedback on how well GPT-3.5 performs in both generating and evaluating a Tell Tale game story.

5 User Evaluation

We enlisted 10 human testers to play a game of Tell Tale using our web application, where each participant played the game individually along with the computer. One important request we gave the human testers was that the human testers should only use three predetermined literary devices: conflict, foreshadowing, and imagery. In this way, we could still test GPT-3.5's ability to include literary devices while also increasing the level of consistency amongst the human testers, especially given that each human tester will be using a different random set of Tell Tale cards.

After testing a round of Tell Tale on the web application, the human testers were surveyed on the abilities of GPT-3.5. First, the testers were asked "On a scale of 1 to 10, how would you score GPT's story based on coherency and how it includes the selected story and literary elements? (1 being the worst and 10 being the best.)" and "Please explain your reasoning for why you gave GPT's story the score you gave." This allowed the testers to provide us with both quantitative and qualitative feedback. The quantitative feedback would be helpful in comparing how the human testers evaluated GPT-3.5's story versus how GPT-3.5 itself would evaluate its own story. On the other hand, the qualitative feedback would provide a more detailed assessment of the quality of GPT-3.5's story, including on the story's strong points and where the story could be improved. Such feedback would be crucial to improve the prompts provided to GPT-3.5 in the future in order to generate better quality stories.

Second, the testers were asked "Review GPT's evaluation of your story. Evaluate GPT's evaluation. In other words, do you agree or disagree with GPT's evaluation and score of your story and why? How well do you think GPT evaluated your story?" and "Score GPT's evaluation of your story on a scale from 1 to 10 (1 being the worst and 10 being the best)." Like the previous two questions, this provides us with both a quantitative and a qualitative evaluation of GPT-3.5's performance as a Tell Tale story evaluator. Similar questions are asked to the human testers about GPT-3.5's evaluation of its own story.

6 Results

Our preliminary results show that, in general, the human testers had a positive assessment of the performance of GPT-3.5 as a participant in a Tell Tale game. The average quantitative score (out of 10) for the quality of GPT-3.5's story was 6.7. The human testers' ratings for GPT-3.5's story evaluation ability were roughly the same, giving GPT-3.5's evaluation of the human testers' stories an average score of 7.0 and giving GPT-3.5's evaluation of its own stories an average score of 7.2.

The qualitative feedback provided by the human testers also supports a positive appraisal of GPT-3.5's storytelling abilities, with some of the human testers describing how GPT-3.5 "incorporated all the elements of the story that it needed to" and how GPT-3.5's story "was very easy to read, engaging, and

included the required literary elements." Of course, not all the feedback was universally positive, with one human tester noting that "the story felt 'forced', with GPT trying to be too literal in its use of the story card items." Another tester was harsher, noting how "The imagery is detailed but predictable and schematic, so that it brings to mind a cartoony picture rather than engaging the imagination in filling out and entering a scene."

Most human testers also generally praised GPT-3.5's evaluation ability, with some testers noting how quickly GPT-3.5 was able to evaluate their stories and recognize that the story included certain literary features ("I was surprised it recognized my foreshadowing in the last sentence" notes one tester). Nevertheless, GPT-3.5's evaluation ability was not perfect. For instance, one tester felt as though GPT-3.5 should have been able to detect the weakness in a certain story's plot.

7 Conclusions

In our initial testing and in the evaluation of our web application, we found that GPT-3.5 was proficient in generating coherent narratives and generating a meaningful interpretation of a set of initial observations (Tell Tale cards, in this research). Such proficiency is evidenced by the generally high scores human testers have given to many of GPT-3.5's stories. In addition, GPT-3.5 has proven adept at evaluating other Tell Tale stories too, with the human testers giving GPT-3.5's story evaluations high scores as well.

Our research also highlights several areas where GPT-3.5's game-playing abilities could be improved, either through more refined prompts or through improvements to the language model itself. Some of the human testers wrote in their qualitative assessments their opinion that some of GPT-3.5's stories were somewhat contrived and their inclusion of Tell Tale cards a bit forced. In our opinion, this research is a step towards enabling large language models to perform complex and interactive narrative tasks.

References

1. Elson, D.: DramaBank: Annotating agency in narrative discourse. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 2813–2819. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012)
2. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., Leahy, C.: The Pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020)
3. Gordon, A.S., Spierling, U.: Playing Story Creation Games with Logical Abduction. In: International Conference on Interactive Digital Storytelling. pp. 478–482. Springer (2018)
4. Gordon, A.S., Wang, T.S.: Narrative text generation from abductive interpretations using axiom-specific templates. In: Mitchell, A., Vosmeer, M. (eds.) Interactive Storytelling. pp. 71–79. Springer International Publishing (2021)

5. Han, R., Chen, H., Tian, Y., Peng, N.: Go back in time: Generating flashbacks in stories with event temporal prompts. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States (Jul 2022)
6. MosaicML NLP Team: Introducing MPT-7B: A new standard for open-source, commercially usable LLMs (2023), www.mosaicml.com/blog/mpt-7b, accessed: 2023-08-01
7. Peng, N., Ghazvininejad, M., May, J., Knight, K.: Towards controllable story generation. In: Proceedings of the First Workshop on Storytelling. pp. 43–49. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/W18-1505, https://aclanthology.org/W18-1505
8. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018), https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035
9. Roemmele, M.: Identifying sensible lexical relations in generated stories. In: Proceedings of the First Workshop on Narrative Understanding. pp. 44–52. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/W19-2406, https://aclanthology.org/W19-2406
10. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023)
11. Trabasso, T., van den Broek, P.: Causal thinking and the representation of narrative events. Journal of Memory and Language **24**(5), 612–630 (1985). https://doi.org/https://doi.org/10.1016/0749-596X(85)90049-X