More human than human: LLM-generated narratives outperform human-LLM interleaved narratives

Zoie Zhao, Sophie Song*
Department of Computer Science, Smith College
Northhampton, MA, USA
(zzhao39,ssong25)@smith.edu

Bridget Duah, Jamie C. Macbeth
Department of Computer Science, Smith College
Northhampton, MA, USA
(bduah,jmacbeth)@smith.edu

Scott Carter, Monica Van, Nayeli Bravo, Matthew Klenk, Katherine Sieck, Alexandre Filipowicz
Toyota Research Institute
Los Altos, CA, USA
(scott.carter,monica.van.ctr, nayeli.bravo.ctr,matt.klenk, kate.sieck,alex.filipowicz)@tri.global

ABSTRACT

Narrative story generation has gained emerging interest in the field of large language models. The present paper aims to compare stories generated by an LLM only (non-interleaved) with those generated by interleaving human-generated and LLM-generated text (interleaved). The study's hypothesis is that interleaved stories would perform better than non-interleaved stories. To verify this hypothesis, we conducted two tests with roughly 500 participants each. Participants were asked to rate stories of each type, including an overall score or preference and four facets—logical soundness, plausibility, understandability, and novelty. Our findings indicate that interleaved stories were in fact less preferred than non-interleaved stories. The result has implications for the design and implementation of our story generators. This study contributes new insights into the potential uses and restrictions of interleaved and non-interleaved systems regarding generating narrative stories, which may help to improve the performance of such story generators.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; Machine learning; • **Human-centered computing** → *HCI design and evaluation methods.*

KEYWORDS

Story generator, neural networks, gaze detection, text tagging

ACM Reference Format:
Zoie Zhao, Sophie Song, Bridget Duah, Jamie C. Macbeth, and Scott Carter, Monica Van, Nayeli Bravo, Matthew Klenk, Katherine Sieck, Alexandre Filipowicz. 2023. More human than human: LLM-generated narratives outperform human-LLM interleaved narratives. In *Creativity and Cognition (C&C '23), June 19–21, 2023, Virtual Event, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3591196.3596612

*Both authors contributed equally to this research.

C&C '23, June 19–21, 2023, Virtual Event, USA

ACM ISBN 979-8-4007-0180-1/23/06.
https://doi.org/10.1145/3591196.3596612

1 INTRODUCTION

Large Language Models (LLM) such as GPT-3+ have brought about a revolution in natural language processing and have expanded the possibilities of automated text generation. By using machine learning algorithms, these models evaluate extensive amounts of language data in order to develop the capacity to generate sentences that are indistinguishable from those written by humans. One specific application of this technology is the generation of narrative stories, which has the potential to transform the area of creative writing and open up new opportunities for storytelling in a variety of settings. Nevertheless, the quality of stories generated by LLM varies widely, and our understanding of how individuals perceive them is limited.

In this paper, we present a study that compares stories generated by an LLM only (non-interleaved) with those generated by interleaving human-generated and LLM-generated text (interleaved).

2 RELATED WORK

Alabdulkarim et al. investigated how to improve a story generator, including the controllability of story generation, acquiring commonsense knowledge, frameworks of stories, and other challenges such as creativity [1]. Fan et al. researched hierarchical neural story generation using a fusion mechanism and evaluated stories generated by this hierarchical structure [3]. They evaluated with automatic evaluation, which measured model perplexity and prompt ranking accuracy, and human evaluation, which tested the pairing of prompt and stories. They concluded that generating with a hierarchy frame performed better in terms of fluency, topicality, and overall quality.

In our own prior work, we focused on how to maintain the coherence of narrative stories generated by GPT [4]. This study did not include an evaluation of interleaved and non-interleaved stories; this paper is a continuation of the prior work focusing on the evaluation of the two approaches.

3 SYSTEM DESCRIPTION

We tested non-interleaved stories generated directly from GPT (we used GPT-3/3.5 for these studies) against interleaved stories, which alternate between GPT and a human-generated story at each step (our system also supports stories generated by any computational method, such as TALE-SPIN [4]). For interleaved stories, GPT

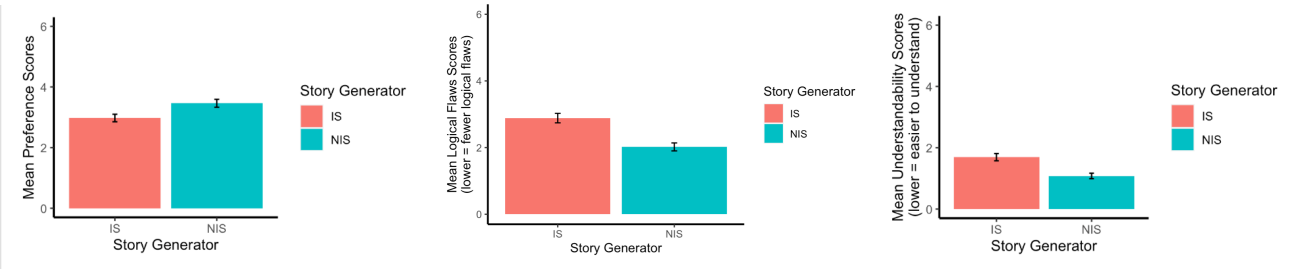

Figure 1: Bar plots comparing the Preference Score (left), Logical Flaws (middle), and Understandability (right) of Interleaved Stories (IS) and Non-Interleaved Stories (NIS).

receives one sentence from the human-generated story prompt, responds to it, receives the next prompt sentence, and continues until it reaches the end. The non-interleaved stories received the entire human-generated story as the lone prompt for the LLM system.

We hypothesize that an interleaved system generates better and more human-like text than a non-interleaved system. Specifically, since in the interleaved system, GPT always responds to the previous sentence in the story frame, the generated story overall should be tighter than the output from the non-interleaved system. We conducted two types of evaluations: a single evaluation of each story type as well as a pairwise evaluation comparing story types. We collected participants' responses and preferences and analyze them with R Studio.

4 METHOD

We generated 10 interleaved and 10 non-interleaved stories for topics related to EV cars, beaches, and cooking. We designed two 20-min surveys using Qualtrics, an absolute Likert scale survey evaluating each story independently, and a pairwise comparison survey, to record participants' preferences on the generated stories.

For the independent-story survey, 477 participants were shown 6 random stories from 60 samples to read and rate. In each story, they were asked to answer the extent to which they liked the story, and the details regarding four facets: logical flaws, plausibility, understandability, and novelty of the stories (7-point scales). For the pairwise survey, 502 participants were shown a pair of interleaved and non-interleaved stories of each of the 3 topics and were asked to select a better story from each pair. Questions about the four facets were also prepared to choose a better-fitting story for each criterion. After completing the ratings, participants answered a series of demographic questions and an open-ended question about their choice process.

5 RESULTS

5.1 Absolute Survey

We first measured the participants' story generation preferences using a logistic mixed effects model (LME) implemented with the lme4 library in R [2]. This model sets the average ratings of interleaved stories as a baseline reference to estimate how different those of the non-interleaved stories are. Overall, participants tended to prefer non-interleaved stories, responding 0.49 Likert points higher than interleaved stories (p<2e-16). Specifically, participants rated non-interleaved stories as having fewer logical flaws than interleaved stories (Estimate=-0.87, p<2e-16), and were easier to understand (Estimate=-0.63, p<2e-16). However, technologies described in interleaved stories seemed more plausible in the not-too-distant future than that in non-interleaved stories (Estimate=-0.145, p=0.0004). There was no statistically significant difference in the novelty scores.

5.2 Pairwise Survey

We used an exact binomial test to measure whether the ratio of the participants who selected interleaved stories over non-interleaved is significantly different for the preference and each of the other four dimensions. The result indicated that people prefer stories generated by non-interleaved systems more than interleaved (CI=[0.397-0.447], p=1.30e-09), with non-interleaved having fewer logical flaws (CI=[0.386-0.436], p=5.27e-12), seeming more plausible in the near future (CI=[0.440-0.491], p=0.009), and being more understandable (CI=[0.383-0.433], p=6.04e-13). However, we found that participants considered interleaved stories to be more novel than the non-interleaved stories (CI=[0.541-0.592], p=3.70e-07).

Furthermore, running a logistic mixed effects regression showed that participants reading the stories about beaches were more likely to prefer a non-interleaved system than those who read about car stories (Estimate=-0.36, z=-2.735, p=0.00624).

6 DISCUSSION

Contrary to our hypothesis, the results of this study indicate that participants preferred the non-interleaved stories over the interleaved story.

One possible explanation for this result is the vulnerability of the setup of the interleaved system. Our interleaving function is only based on the end of GPT instead of both ends, hence when interleaving, the interleaved system does not view the response from GPT nor adjust its prompt sentence based on the generated response. Therefore, it is possible that GPT responds with one sentence and the next human-generated sentence contains the same piece of information. This creates a redundant reiteration with an unnatural sentence connection that a lot of the readers noticed.

7 CONCLUSION AND FUTURE WORK

In this paper, we conducted two surveys to study and evaluate the text generated by GPT. Our results showed that people found the

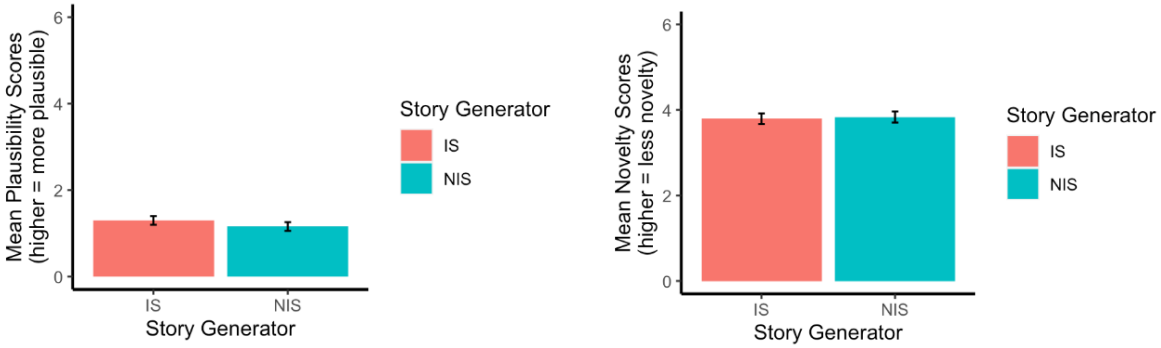

Figure 2: Bar plots comparing the Plausibility (left) and Novelty (right) of Interleaved Stories (IS) and Non-Interleaved Stories (NIS).

non-interleaved stories better than the interleaved stories in general with fewer logical errors and easier understandability. In future studies, we will focus on building a better interleaving system that better integrates and adapts human and GPT responses.

ACKNOWLEDGMENTS

Thank you to Toyota Research Institute for funding this work.

REFERENCES

[1] A. Alabdulkarim, S. Li, and X. Peng. 2021. Automatic Story Generation: Challenges and Attempts. In *Proceedings of the Third Workshop on Narrative Understanding*. 72–83.

[2] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).

[3] A. Fan, M. Lewis, and Y. Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 889–898.

[4] J. Xiang, Z. Zhao, M. Zhou, M. McKenzie, A. Kilayko, J. C. Macbeth, S. Carter, K. Sieck, and M. Klenk. 2022. Interleaving a Symbolic Story Generator with a Neural Network-Based Large Language Model. In *Proceedings of the Tenth Annual Conference on Advances in Cognitive Systems*. The Cognitive Systems Foundation, Arlington, VA.