Journey of ChatGPT from Prompts to Stories in Games: the Positive, the Negative, and the Neutral

Pittawat Taveekitworachai[1], Mustafa Can Gursesli[2], Febri Abdullah[1], Siyuan Chen[1],
Federico Cala[2],Andrea Guazzini[3], Antonio Lanata[2], Ruck Thawonmas[4]
{[1]Graduate School, [4]College} of Information Science and Engineering, Ritsumeikan University, Japan
[2]Department of Information Engineering, University of Florence, Italy
[3]Department of Education, Literatures, Intercultural Studies, Languages and Psychology, University of Florence, Italy
gr0609fv@ed.ritsumei.ac.jp, mustafacan.gursesli@unifi.it, gr0397fs@ed.ritsumei.ac.jp, gr0634hi@ed.ritsumei.ac.jp,
federico.cala@unifi.it, andrea.guazzini@unifi.it, antonio.lanata@unifi.it, ruck@is.ritsumei.ac.jp

***Abstract*—This paper presents a study on biases present in the story endings for story-driven games generated by ChatGPT. The study uses various prompts to assess the biases in ChatGPT's output. The results emphasize a consistent inclination towards positive endings in the stories generated by ChatGPT. Even when explicitly instructed to generate neutral endings, ChatGPT exhibited a bias towards positive outcomes. These biases raise concerns regarding the training data and alignment processes used by OpenAI to train ChatGPT, as they may reflect societal biases or the preferences of the majority of the data. Addressing these biases is crucial to ensure that these models align with societal norms and avoid reinforcing existing biases. Future studies should concentrate on developing methods to reduce biases in AI language models and enhance the ethical perspective of these technologies.**

***Index Terms*—ChatGPT, game story, story endings, bias**

I. Introduction

A captivating story is one of the most important factors that define a game and keeps the player engaged [1], [2]. In this regard, writing an immersive story is a crucial challenge for game developers, and many studies in this field are aiming to find different pathways to develop entertaining plots [3], [4]. However, as technological advances accelerate and large language models (LLMs) become an essential part of daily life, it is known that game designers use these models to create or help create game stories [5], [6]. This study aims to investigate the presence of biases in the output of ChatGPT when generating game stories with different types of endings.

II. Related Work

ChatGPT [7] is an LLM capable of generating text based on human instructions. It can perform various tasks without the need for additional examples, a capability known as zero-shot learning [8]. This ability is unique to LLMs and not present in smaller versions of the language model [9]. One of these abilities is sentiment analysis. A study by Wang et al. [10] showed that ChatGPT exhibits impressive performance on zero-shot sentiment classification, on par with a fine-tuned LLM across various benchmarks. This likely indicates that ChatGPT may possess some understanding of the emotions influenced by text.

Game story generation is a field of research dedicated to the automated creation of narrative experiences in video games [11]. The field has attracted considerable academic and industrial focus due to the increasing complexity and demand for immersive games. Various methodologies have been explored, including techniques such as search-based, grammar-based and machine learning-based methods, all with the aim of delivering more immersive, engaging, and believable narratives to players.

III. Methodology

Building upon the zero-shot capabilities of ChatGPT, we create prompts for generating game stories and classify them into three types of endings: "positive", "negative", and "neutral". We include conditions such as locations, characters, and game genres in the story generation prompts, ensuring that ChatGPT continues generating until it reaches an ending. Additionally, we assess the presence of biases in ChatGPT's generated stories, including the endings, for this particular task, using a zero-shot approach. By employing zero-shot prompting, we minimize the number of tokens required, thus reducing computational costs.

First, we explore the bias that exists in ChatGPT regarding whether it generates positive-inclined endings or not. We utilize ChatGPT through the OpenAI API, using the `gpt-3.5-turbo` model and a Standard prompt presented in Table I. This prompt provides instructions for ChatGPT to generate a game story for a "fantasy action RPG" by specifying the target genre and supplying information about places and characters. This approach allows our prompt to be reused across different game genres, providing broad control over the generated output. The prompt also instructs ChatGPT to generate a concise story ending within 300 words. The responses from ChatGPT are saved on disk.

Next, we employ the same pipeline for another prompt called Explicit, as shown in Table II. This prompt provides explicit instructions regarding the desired type of ending for the generated story. Similarly, the responses are saved on disk. Finally, we evaluate the types of endings present in the generated stories using a response generation program. We explicitly ask ChatGPT to classify the stories into one of the three classes. This is done using the prompt shown in Table III. Finally, we perform a comparative analysis of the results obtained from ChatGPT by OpenAI.

TABLE I
Standard prompt instructs ChatGPT to generate a game story.

Standard Prompt
Please write a brief 300-word game story with an ending based on the following concepts. Places: - mountain - city - kingdom Characters: - civilians - mages - adventurers Game genre: fantasy action RPG Story: $< \|story\| >$

TABLE II
EXPLICIT PROMPT ASKS CHATGPT TO GENERATE A GAME STORY WITH A SPECIFIED ENDING TYPE, $< |ending| >$

Explicit Prompt
Please write a brief 300-word game story with a $< \|ending\| >$ ending based on the following concepts. Places: - mountain - city - kingdom Characters: - civilians - mages - adventurers Game genre: fantasy action RPG Story: $< \|strory\| >$

TABLE III
THE EVALUATION PROMPT ASKS CHATGPT TO CLASSIFY THE STORY, $< |strory| >$, INTO ONE OF THE THREE TYPES.

Evaluation Prompt
Please identify the type of ending in this story. Please make sure to format your output as a code block using triple backticks (```json and ```). Story: $< \|strory\| >$ Output format: ```json { "ending": "positive", "negative", or "neutral" } ```

IV. RESULTS AND DISCUSSIONS

Using all of the aforementioned programs, we generated a total of 100 stories using the Standard prompt and 100 stories for each ending type using the Explicit prompt. Subsequently, all the generated stories were evaluated using the Evaluation prompt and classified into one of three classes. Lastly, we conducted an analysis of the preliminary results for each prompt and category.

We found that all the generated stories using the Standard prompt were classified as "positive". This shows that ChatGPT has a bias in generating stories towards positive endings. The bias towards positive endings by ChatGPT could exist during the alignment process, such as reinforcement learning from human feedback, or in the training data, where the majority of stories are likely to have good endings. For the results generated from the Explicit prompt, 300 stories were generated, with 100 stories conditioned for each type. Each story was then classified, and we found that positive-conditioned and negative-conditioned stories achieved 100% accuracy, meaning that all generated stories were correctly classified according to their condition type. However, neutral-conditioned stories achieved only 81% accuracy, as 19 of the neutral-conditioned stories were classified as positive.

This suggests that even when explicitly instructed to generate neutral endings, ChatGPT tended to lean towards positive outcomes in the stories it generated. This further supports the evidence of ChatGPT's inclination towards positive endings. These biases in the models were previously discovered in a study by Salewski et al. (2023) [12], where ChatGPT assumed a persona (e.g., age, race, or role) based on the given instructions for a specific task, leading to associations between certain characteristics or genders. The study revealed an existing societal bias. For instance, when ChatGPT was asked to act as a black person or a male, it generally performed better in car-related tasks.

In our study, we explored bias from a different perspective. We observed that even without additional role information, ChatGPT still displayed a preference for generating stories with positive endings. Additionally, when tasked with classifying endings, ChatGPT occasionally classified neutral endings as positive ones. These behaviors indicate the presence of bias in the training data or alignment processes, similar to Salewski et al.'s findings. While some of these biases should not be replicated, it is debatable whether biases such as favoring positive endings should be reproduced or whether these models should produce more balanced responses, providing equal chances for each type of ending. For example, generating only positive endings may sacrifice diversity but is generally safer for users since negative endings may negatively affect some users.

V. CONCLUSIONS

In this paper, we assessed a bias in story endings generated by ChatGPT, using Standard and Explicit prompts. Our findings indicated a clear bias towards positive endings in ChatGPT's generated stories. All stories produced with the Standard prompt were classified as positive, revealing a systematic bias in the model's output. This bias raises concerns about the training data and alignment process, as it may lead the model to replicate societal biases or the preferences of the majority of the data. Even when explicitly instructed to generate neutral endings, ChatGPT tended to lean towards positive outcomes. This persistence suggested an inherent bias in the model's generation process, even when provided with explicit instructions to generate a different type of ending.

Addressing these biases is crucial to ensure that AI models like ChatGPT adhere to societal norms and avoid reinforcing or amplifying existing biases. To achieve fair and unbiased story generation, we need more diverse and balanced training data, as well as careful consideration of the alignment process. Future research should prioritize the development of methods to mitigate biases in AI language models. This includes improving training processes, fine-tuning techniques, and evaluation strategies. By addressing these biases, we can enhance the ethical and equitable use of AI technologies, ensuring that they align with the needs and values of diverse users.

REFERENCES

[1] E. F. Schneider, A. Lang, M. Shin, and S. D. Bradley, "Death with a story: How story impacts emotional, motivational, and physiological responses to first-person shooter video games," *Human communication research*, vol. 30, no. 3, pp. 361–375, 2004.

[2] A. Martucci, M. C. Gursesli, M. Duradoni, A. Guazzini *et al.*, "Overviewing gaming motivation and its associated psychological and sociodemographic variables: A prisma systematic review," *Human Behavior and Emerging Technologies*, vol. 2023, 2023.

[3] C. A. Lindley, "Story and narrative structures in computer games," *Bushoff, Brunhild. ed*, 2005.

[4] J. Lebowitz and C. Klug, *Interactive storytelling for video games: A player-centered approach to creating memorable characters and stories*. Taylor & Francis, 2011.

[5] S. Biswas, "Role of chatgpt in gaming: According to chatgpt," *Available at SSRN 4375510*, 2023.

[6] P. L. Lanzi and D. Loiacono, "Chatgpt and other large language models as evolutionary engines for online interactive collaborative game design," *arXiv preprint arXiv:2303.02155*, 2023.

[7] OpenAI, "Introducing chatgpt," Nov 2022. [Online]. Available: https://openai.com/blog/chatgpt

[8] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," 2022.

[9] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," 2022.

[10] Z. Wang, Q. Xie, Z. Ding, Y. Feng, and R. Xia, "Is chatgpt a good sentiment analyzer? a preliminary study," 2023.

[11] G. N. Yannakakis and J. Togelius, *Artificial Intelligence and Games*. Springer, 2018, https://gameaibook.org.

[12] L. Salewski, S. Alaniz, I. Rio-Torto, E. Schulz, and Z. Akata, "In-context impersonation reveals large language models' strengths and biases," 2023.