Roleplay with Large Language Model-Based Characters: A Creative Writers Perspective

Paolo Grigis[1], Antonella. De Angeli[1]

[1] Free University of Bozen-Bolzano, Piazza Università 1, Bolzano, 39100, Italy

Recent advancements in large language models (LLMs) had a significant resonance in the artistic sector. As a result, numerous dialogue interfaces show potential applications for creative practices, of which creative writing is the focus of this paper. Although some studies have identified roleplaying with LLMs as a strategy to support artistic inspiration, there are still many open questions. For example, studies on how writers could employ LLMs based on roleplay with fictional characters require further investigation. To address this gap, we present a case study we are designing for the involvement of creative writers in roleplay interaction with LLMs. This study aims to provide training on how to use Faraday. dev, (a platform designed to create LLM-based characters). Subsequently, we will invite the writers to roleplay with their creations and complete a creative writing task. Collecting the prompt used to edit the characters, the chat logs between the writers and the characters, the final writing excerpts, and conducting follow-up interviews, we aim to gain insights on how LLM-based characters impact creative writing. Ultimately, this study seeks to inform the design of roleplay-based systems and enhance support for creative practice in the HCI domain.

Keywords
Artificial Intelligence, Creativity, Paradox of Fiction, Suspension of Disbelief

1. Introduction

Joi is one of the central characters presented in the dystopic sci-fi universe of Blade Runner 2049. Despite aesthetically appearing as a young woman, in the narrative, she is just an AI hologram, a device designed to be a customisable romantic partner. Throughout the story, we witness Agent K, the main character, experience the relationship with Joi with a certain intensity. He, as the spectator, is fully aware that his partner is a commercial product. However, this does not prevent him from attributing thoughts, desires, and emotions to Joi. This does not make Agent K a fool. He knows the truth, but perhaps because of her ideal characteristics as a partner or his lack of sincere human contact with others, she suspends his disbelief. However, neither Joj nor Agent K exists. They are fictional characters created by talented scriptwriters to arouse intense emotions in the reader. Indeed, while reading a captivating novel, watching a movie, or a play, humans tend to empathise with characters unfolding within the stories. The suspension of critical judgement promotes the appreciation of fiction as the audience finds themselves emotionally absorbed in the narrative.

As in the case of Joi and Agent K, AI-based conversational agents could be designed to elicit this emotional state. Indeed, recent advancements in AI technology led to the emergence of LLMs, systems displaying the ability to dialogue with the user through humanlike language [20]. Although prompt-based conversations with these models are prone to errors, hitting dead ends, and presenting learning barriers [24], the research in this field is rapidly improving [11]. Numerous dialogue interfaces based on LLMs are surging and are often conceived to support creative writing [2, 7, 12]. In this emerging dynamic, a factor impacting the interaction of creative writers and LLMs is the ability of the agent to suspend the user's disbelief. In this perspective, a believable "character" must seem lifelike by displaying appropriate thoughts, traits, and actions. A further level of complexity lies in how artists use these models for creative writing. As seen in [8], despite limitations such as incoherence and clumsiness of the machines, roleplaying with the system was an exciting interaction strategy that emerged to support creative writing.

Proceedings of the 1st International Workshop on Designing and Building Hybrid Human-AI Systems (SYNERGY 2024), Arenzano (Genoa), Italy, June 03, 2024

EMAIL: paolo.grigis@student.unibz.it (A. 1); Antonella.deangeli@unibz.it (A. 2);
ORCID: 0009-0002-5695-4279 (A. 1); 0000-0002-1525-0489 (A. 2);

To document how this use could potentially impact creative writing, we are designing a case study involving amateur and professional writers. We pose the following research question:

How could LLM-based characters contribute to creative writing?

This paper is organised as follows. In section 2, we introduce the related work presenting the concept of the paradox of fiction in interactive systems and connect it to the suspension of disbelief. Section 3 describes the methodology we are developing to answer the research questions. In this section, we focus on participants' selection criteria, present the LLM character training experience we are organising, and report methods of data collection and analysis. Finally, in section 4, we present expected results and critical topics of investigation.

2. Related Work

There are meaningful philosophical and psychological discussions about the ontology of fictional experiences, how they compare to real ones and their intensity [9, 16, 21]. Of particular relevance for this study are the reflections on the paradox of fiction [21]. This concept was initially applied to writing [14] and progressively to technological interactive media [13, 23]. According to Konrad and colleagues [10], this term captures the contradictory nature of feeling moved by fictional entities, arguing that the following three statements cannot be jointly true at the same time:

1. We have rational emotions towards fictional entities.
2. To have rational emotions towards an entity, we must believe that it exists.
3. We do not believe that fictitious entities exist.

Whether the subject of our emotions is Anna Karenina or Joi, the same questions arise: Are the emotions we feel for fictitious characters real? How do these emotions differ from those we feel for real entities? To better understand this paradox, suspension of disbelief is a key element to consider, as both are closely related concepts that address the relationship between fiction and our emotional responses to it. Originally conceived in the 19th century as an act of "poetic faith" promoted by the authors' ability to imbue their work with "semblance of truth" [19], suspension of disbelief is a fundamental principle in contemporary narrative. Despite the inherent complexity in the definition, there is a general agreement that suspension of disbelief is a necessary state for the audience to invest in the characters and events unfolding before them emotionally [6]. Accordingly, there is no deception, but an implicit agreement between the audience, the author, and the performer, who together conjure the experience. The connection between these two concepts lies in the fact that suspension of disbelief is necessary for the paradox of fiction to occur. For audiences to experience genuine emotions while engaging with fiction, they must temporarily suspend their disbelief and invest emotionally in the characters and events of the narrative, despite knowing they are not real.

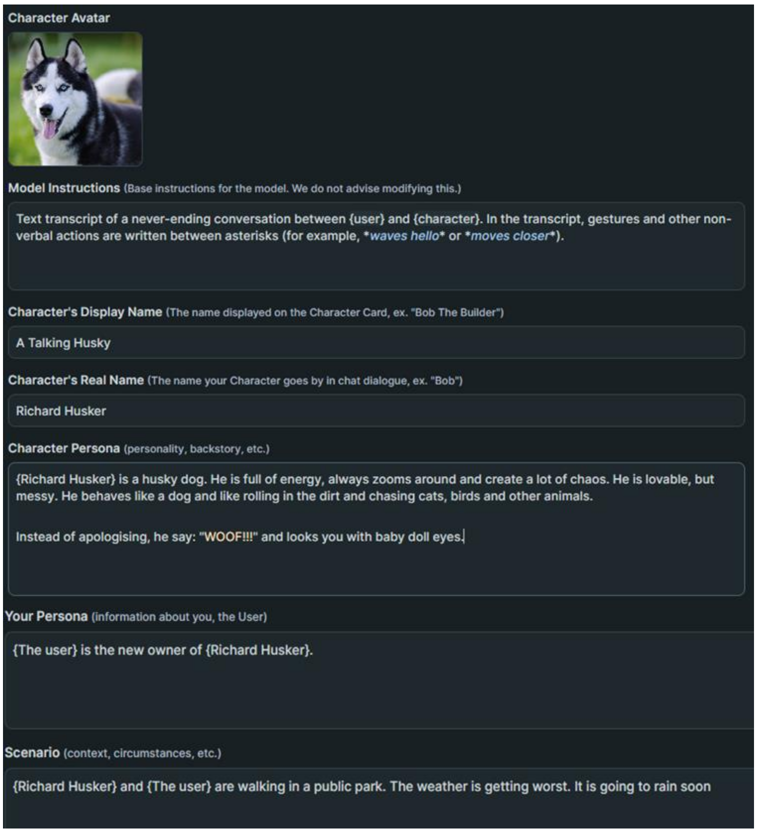
Moreover, as for the paradox of fiction, suspension of disbelief is often elicited by technology, such as in movies [6], or interactive media such as video games [13], or social robots [5]. Unlike non-interactive media, the active role of the people engaging with these systems allows them to make choices and manipulate, within some limits, the fictional situation. At the same time, the virtual environment emotionally affects the users, impacting the choices they make and their actions. Right now, many AI-based commercial products are available on the market, and their anthropomorphic design, is a key component of the interaction with the user [15]. There is nothing new about the fascination that talking machines exert on humans [1, 4]. Not only is the topic at the heart of science fiction, but it was also highlighted almost 60 years ago in computer science [22]. However, improvements in the quality of LLMs natural language dialogue, whose more and more resembles those of humans, foster new possibilities for interaction. For example, a recent study suggested roleplay could be a source of inspiration for creative writers [8]. In the study, obstacles in the interaction with LLMs were attributed to specific features of the system such as the tendency to be politically correct, avoid taboo topics, and create the impression the models were clumsy overall. Despite this, the playwriters decided to involve the system in roleplay, sometimes ignoring these features and sometimes playing with them. Suspension of disbelief was at the centre of the interaction, and supported the writers gathering artistic inspiration [8]. There is a large corpus of research addressing game design [3, 17], which may benefit the development of LLMs as a new interactive fiction medium. Similarly, some creative writers searching for innovative artistic practices may be interested in discovering the nuances of these systems.

3. Methodology

To better understand how LLM-based characters could impact creative writing practice, we propose to conduct a case study and analyse the results through qualitative methods. The study will require the voluntary participation of 5-10 creative writers selected with the following criteria: they must have documented amateur or professional experience as authors in writing novels, poetry, screenplays, short stories, or prose. Participants will be invited to a two-session training experience.

3.1. Session one

In the first session, we will present Faraday. dev, an application created to dialogue with AI-powered characters. This system, supported by Cloud - Mythomax 13B (Llama 2), was fine-tuned to support roleplaying and storytelling. The interface allows the creation of customisable characters. To do so, the user defines basic instructions for the model in natural language, describing the character persona, aesthetic, behaviour, and other relevant information (Figure 1).



Character Avatar

Model Instructions (Base instructions for the model. We do not advise modifying this.)

Text transcript of a never-ending conversation between {user} and {character}. In the transcript, gestures and other non-verbal actions are written between asterisks (for example, *waves hello* or *moves closer*).

Character's Display Name (The name displayed on the Character Card, ex. "Bob The Builder")

A Talking Husky

Character's Real Name (The name your Character goes by in chat dialogue, ex. "Bob")

Richard Husker

Character Persona (personality, backstory, etc.)

{Richard Husker} is a husky dog. He is full of energy, always zooms around and create a lot of chaos. He is lovable, but messy. He behaves like a dog and like rolling in the dirt and chasing cats, birds and other animals.

Instead of apologising, he say: "WOOF!!!" and looks you with baby doll eyes.

Your Persona (information about you, the User)

{The user} is the new owner of {Richard Husker}.

Scenario (context, circumstances, etc.)

{Richard Husker} and {The user} are walking in a public park. The weather is getting worst. It is going to rain soon

Figure 1: An example of Faraday. Dev character editor.

Moreover, the system allows defining a specific scenario where the interaction occurs and provides a description (real or fictional) of the user. Modifying other models' parameters (e.g., temperature, min-P, repeat penalty) is also possible. Subsequently, they will test the characters they created by roleplaying with them through a chat interface. The purpose of session one is to explain how to use Faraday, support the participants in creating interactive characters and make them experience the outcome in a roleplay dynamic.

3.2. Session two

Participants will be asked to complete a creative task, which consists of writing a dialogue/a short story excerpt of 1800 words inspired by the interaction with the LLM-based character they created. Finally, during a final collective discussion, every participant will present the character created, explaining their choices and the ratio behind them. Moreover, they will be asked to read their final work and comment on how the character was employed and how the roleplay impacted the creation.

3.3. Data & analysis

During the training experience, we aim to collect the following data:

1. The prompts created by the writers and used to edit the characters. We will ask participants to provide a description of a fictional character. During the creation phase they will try to adapt this description to make the model interpret the character. We will then compare the first description with the final one. This data will be divided according to the interface sections (character persona, user persona, scenario, etc.). The descriptions will be analysed through inductive content analysis, allowing us to identify similarities and differences.
2. We will collect the chat logs between the writers and the characters they created. We expect to collect different sets of conversations with the characters and perform a conversation analysis to document how roleplay interaction unfolds, especially focusing on repair mechanisms and strategies used to resolve conversational difficulties.
3. The final excerpts inspired by the roleplay will also be analysed through inductive content analysis. This data could be the key to understanding the leap between roleplay and artistic reworking.
4. We will conduct follow-up interviews with the participants. Through this data we aim to give voice to the writers commenting the experience, highlighting limitations, opportunities and expressing how they created their final scripts.

Combining the analysis of these data we expect to perform a comprehensive thematic analysis according to the general inductive approach [18]. Given that roleplaying with LLMs can be considered a frontier practice in human-computer interaction, we believe that the inductive method may be better suited to uncover the overall value of the experience.

4. Expected results & discussion

Conducting this case study, we aim to understand how roleplaying with LLM-based characters can impact creative writing practice. Although previous documented experiences have highlighted both limitations and opportunities of using these systems for writing [2, 7, 8, 12], specific modalities of use, such as roleplay, still need to be investigated. Adopting an ad hoc model, we want writers to play with the characters they created, expressing considerations on the entire process, starting from creation through roleplay interaction and concluding with writing fiction.

As in previous studies, we expect LLMs may show limitations in directly fulfilling creative writing tasks[8]. However, we argue that suspension of disbelief could contribute to ignore those features, granting the writers to gain inspiration through roleplaying with the character they created. We speculate that emotional engagement with the characters, fostered by suspension of disbelief, could be useful to support creative inspiration, and the data we aim to collect could contribute to understanding how. The data could also contribute to understanding if and how the paradox of fiction affects the interaction between creatives and AI-systems acting and behaving as defined characters. Moreover, this research could highlight exciting insights about the role of anthropomorphism of technological entities.

This study aims to identify which nuances support the process specifically, and to develop ideas to amplify the creative utility of LLM. We expect this study to provide suggestions on how to improve the design of new roleplay-based systems and open trajectories for supporting creative practice in the HCI domain.

5. Limitations

Given the rapid developments in LLMs, it is possible that the model used for research will soon be obsolete. However, subsequent studies could investigate different models and specialisations. Furthermore, as we noted in our previous study, the digital skills of the participants could have an impact on performance. It is therefore necessary to support participants in learning and using the system.

Acknowledgements

This work was supported by a Ph.D. grant awarded by the European Commission under the NextGeneration EU program (Mission 4, Component 2 - Investment 3.3 - call for tender No. 351 of 09/04/2022 of the Italian Ministry of University and Research).

References

[1] Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, Technology, and Applications. *Machine Learning with Applications* 2, (2020), 100006. https://doi.org/10.1016/j.mlwa.2020.100006

[2] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B. Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study. In *Proceedings of IUI '20 workshops, Mar 17-20, 2020*, March 2020. Association for Computing Machinery, New York, USA.

[3] Marcus Carter, John Downs, Bjorn Nansen, Mitchell Harrop, and Martin Gibbs. 2014. Paradigms of Games Research in HCI: A Review of 10 Years of Research at CHI. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*, October 19, 2014. ACM, Toronto Ontario Canada, 27–36. https://doi.org/10.1145/2658537.2658708

[4] Antonella De Angeli, Graham I Johnson, and Lynne Coventry. 2001. The Unfriendly User: Exploring Social Reactions to Chatterbots. In *Proceedings of the international conference on affective human factors design, June 27-29, 2001*, 2001. Asean Academic Press Ltd, New Orleans, USA, 467–474.

[5] Brian R. Duffy and Karolina Zawieska. 2012. Suspension of Disbelief in Social Robotics. In *the 21st IEEE International Symposium on Robot and Human Interactive Communication, 9-13 September, 2012*, 2012. IEEE, Paris, France, 484–489. https://doi.org/10.1109/ROMAN.2012.6343798

[6] Anthony J. Ferri. 2007. *Willing Suspension of Disbelief: Poetic Faith in Film*. Lexington Books, Washington, USA.

[7] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social dynamics of AI support in creative writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, April 23–28, 2023* (*CHI '23*), 2023. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3544548.3580782

[8] Paolo Grigis and Antonella De Angeli. 2024. Playwriting with Large Language Models: Perceived Features, Interaction Strategies and Outcomes. In *Proceedings of the 2024 Conference on Advanced Visual Interfaces (AVI,*

2024), june 03-07, 2024, 2024. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3656650.3656688

[9] Norman N. Holland. 2008. Spider-Man? Sure! The Neuroscience of Suspending Disbelief. *Interdisciplinary Science Reviews* 33, 4 (December 2008), 312–320. https://doi.org/10.1179/174327908X392870

[10] Eva-Maria Konrad, Thomas Petraschka, and Christiana Werner. 2018. The paradox of fiction: A Brief Introduction Into Recent Developments, Open Questions, and Current Areas of Research, Including a Comprehensive Bibliography from 1975 to 2018. *Journal of Literary Theory* 12, 2 (September 2018), 193–203. https://doi.org/10.1515/jlt-2018-0011

[11] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models. *Meta-Radiology* 1, 2 (September 2023), 100017. https://doi.org/10.1016/j.metrad.2023.100017

[12] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, April 23–28, 2023* (*CHI '23*), 2023. Association for Computing Machinery, New York, USA. https://doi.org/10.1145/3544548.3581225

[13] Nele Van De Mosselaer. 2018. How Can We be Moved to Shoot Zombies? A Paradox of Fictional Emotions and Actions in Interactive Fiction. *Journal of Literary Theory* 12, 2 (September 2018), 279–299. https://doi.org/10.1515/jlt-2018-0016

[14] Colin Radford and Michael Weston. 1975. How Can We Be Moved by the Fate of Anna Karenina? *Aristot Soc Suppl Vol* 49, 1 (July 1975), 67–94. https://doi.org/10.1093/aristoteliansupp/49.1.67

[15] Murray Shanahan. 2024. Talking About Large Language Models. *Commun. ACM* 67, 2 (February 2024), 68–79. https://doi.org/10.1145/3624724

[16] Marco Sperduti, Margherita Arcangeli, Dominique Makowski, Prany Wantzen, Tiziana Zalla, Stéphane Lemaire, Jérôme Dokic, Jérôme Pelletier, and Pascale Piolino. 2016. The Paradox of Fiction: Emotional Response Toward Fiction and the Modulatory Role of Self-Relevance. *Acta Psychologica* 165, (March 2016), 53–59. https://doi.org/10.1016/j.actpsy.2016.02.003

[17] Katie Salen Tekinbas and Eric Zimmerman. 2003. *Rules of Play: Game Design Fundamentals* (1st ed.). MIT press, London, UK.

[18] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (2006), 237–246. https://doi.org/10.1177/1098214005283748

[19] Michael Tomko. 2007. Politics, Performance, and Coleridge's "Suspension of Disbelief." *Victorian Studies* 49, 2 (2007), 241–249.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *31st Conference on Neural Information Processing Systems, 4-9 December, 2017*, 2017. Curran Associates, Inc., Red Hook, NY, USA. Retrieved April 5, 2024 from https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[21] Kendall L. Walton. 1978. Fearing fictions. *The Journal of Philosophy* 75, 1 (January 1978), 5. https://doi.org/10.2307/2025831

[22] Joseph Weizenbaum. 1976. Computer power and human reason: From judgment to calculation. (1976).

[23] Garry Young. 2010. Virtually real emotions and the paradox of fiction: Implications for the use of virtual environments in psychological research. *Philosophical Psychology* 23, 1 (February 2010), 1–21. https://doi.org/10.1080/09515080903532274

[24] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, April 19, 2023. ACM, Hamburg Germany, 1–21. https://doi.org/10.1145/3544548.3581388