"Once Upon a Time There Was a LLM that could write a story", a Study on Human - AI Interaction Through Text - Based Video Games*

Georgios Doukeris[1,*,†], Mike Preuss[1,†] and Giulio Barbero[1,†]

[1]Leiden University, Rapenburg 70, 2311 EZ Leiden, South Holland, Netherlands

Abstract

This study examines how artificial intelligence (AI) can be used to create stories in text-based video games and how players react to them compared to stories written by humans. Inspired by the Turing test, two experiments are conducted: one to explore players' preferences and ability to distinguish between the two types of stories, and another to study how the complexity of the text affects their judgment. Results suggest that, while many participants struggle to identify the story's author, simpler texts are more often seen as written by humans. Moreover, younger participants are more accepting of AI-generated narratives. The research highlights the potential for AI in storytelling while noting current limitations in AI's creative abilities. Additionally, it suggests that the use of AI in creative and entertainment fields could increase as the technology improves.

Keywords

generative AI, text-based games, game AI, creative intelligence, video games

1. Introduction

In recent years, the advent of foundation models has progressively transformed AI into an everyday tool [1]. A field particularly affected by this development is the gaming industry. AI models are used in many areas within the field, including Procedural Content Generation (PCG), Non-Player Characters (NPCs) behavior, Enhanced Personalization, and others [2]. In this study, we research how artificial intelligence can be incorporated into the narrative design process of text-based games and story-based games and what the effects are on users' experience. In particular, we investigate two main research questions:

1. **How does an AI-generated narrative in a text-based game compare with a human-generated one?** Would players be able to distinguish between the two? In order to find an answer, we explore and apply our method, which is strongly influenced by the Turing test [3]. Specifically, we develop the following sub-research questions to focus our investigation:
   - **What parts of the story do humans focus on to detect if it is generated by humans or AI?** Previous research shows that humans fail to reliably identify AI-generated stories [4]. Nevertheless, numerous related strategies have been developed. We identify and test whether these strategies can make AI-generated text more human-like.
   - **How to make AI-generated stories more human?** As mentioned above, in this research, we first aim to test and investigate the human ability to detect generated content. Subsequently, we apply this knowledge to text generation and purposefully attempt to create content that is more likely to be identified as human.
2. **Would the player like a story generated by an AI?** Previous research suggests that humans tend to be biased against AI [5] [9]. While we plan to test this hypothesis, we aim to analyze it further, exploring variations based on participants' characteristics. Furthermore, we raise the following sub-research questions

Proceedings of AI4HGI '25, the First Workshop on Artificial Intelligence for Human-Game Interaction at the 28th European Conference on Artificial Intelligence (ECAI '25), Bologna, October 25-30, 2025

†These authors contributed equally.

g5000d@gmail.com. (G. Doukeris); m.preuss@liacs.leidenuniv.nl (M. Preuss); g.barbero@liacs.leidenuniv.nl. (G. Barbero)

- **How good are humans at distinguishing if a story is generated by a human or AI?** As a corollary of the research questions above, we collect further data about human effectiveness in identifying AI-generated text
- **Would a human "like" the idea of an AI narrator?** As mentioned earlier, research has shown that humans are not comfortable with the idea of AI in their everyday lives. However, we argue that this opinion might change based on individual characteristics (e.g., age) and the context of application (in our case, entertainment).

As the actual object of study, we use an open-source (human-written) text-based game as a starting point. Then, we generate variations of it using GPT4o [16]. Our methodology involves a first exploratory experiment, with a related analysis of the preliminary results, and, subsequently, a more exploitative experiment to test hypotheses emerging from the previous analysis.

2. Related Works

Specifically, our research makes use of **Large Language Models** (LLM). In the last few years, these AI models have significantly impacted many aspects of modern society. In particular, LLMs have been used to explore new frontiers in areas such as automated content generation, human-computer interaction, and language translation [6]. In this regard, game research has made abundant use of this technology [7]. At the same time, new challenges arise with regard to ethics, authenticity, and general misuse of these new technologies [8].

Another important element of the present research is the aforementioned method, which is inspired by the **Turing Test**. In the original version of the Turing test, a human judge evaluates their text interaction with both a machine and a human. In this context, human judgment is the measure of how human-like an AI is. While the Turing Test has influenced AI development for decades, its relevance today is debated; many argue that it focuses too much on human imitation rather than true understanding or reasoning. This holds especially true for modern AI systems such as LLMs, which excel at generating human-like responses but may lack actual comprehension or reasoning. Overall, the Turing test's main drawback is the tendency to focus more on *deceiving* the human participant than on testing actual artificial reasoning [10] [14].

The usage of games as media for conducting Turing tests has been widely explored. For example, Mourning and Lounsbury use the game *Dance Dance Revolution* to investigate whether an algorithm could generate believable beatmaps (key combinations that players are supposed to follow in the game). Using pre-existing songs as a base, they generate a series of beatmaps and use a Turing test to compare them with human-generated ones [11].

Another example worth mentioning is *"Human or Not? A Gamified Approach to the Turing Test"* [12]. This game includes an environment in which the player interacts with either a human or an AI. This game sets up real-time interactions, requiring users to analyze language patterns, personality quirks, and contextual responses to judge whether they are engaging with a person or a bot. The game's environment is designed to mimic real-life social exchanges, allowing the AI to adopt personas with distinct characteristics, such as making deliberate spelling mistakes or using slang to enhance the illusion of humanness. The tool measures success not just by whether the AI can deceive humans but also by how engaging and convincing the interaction feels. This approach extends the Turing Test concept by introducing a dynamic medium for evaluation, which adds depth to understanding how AI can replicate human-like communication in social contexts.

Moreover, the paper "AI Bots in Video Games: A Study of Social Interaction and Turing Test Metrics" examines AI bots' interaction with human players in multiplayer video games. The study uses principles from the Turing Test to evaluate the bots' ability to blend in and mimic human-like behavior, particularly in social and strategic interactions. By focusing on multiplayer games, where social dynamics play a crucial role, the paper extends the Turing Test from a text-based framework to a real-time, decision-based environment. It explores how effectively AI bots can engage with human players without being detected. This analysis highlights how AI can be designed to pass as humans in environments that

require more than just linguistic skill, challenging the AI to display emotional intelligence, adaptability, and strategic thinking [13].

The examples above illustrate the validity of the Turing test-inspired methods in game research, albeit with their limitations.

3. Relevance

Understanding how AI-generated text compares to human-written narratives is becoming increasingly important. It not only sheds light on what AI is capable of today but also helps us explore the strategies people use to spot AI-generated content. Moreover, by investigating participants' preferences, we aim to reveal how much trust and acceptance there is towards AI-generated content in game contexts. These are factors that will likely impact how AI is integrated into creative and educational spaces in the future.

The research also extends into the psychological and perceptual aspects of how people interact with AI-created stories. Biases *for* or *against* AI-authored content influence how narrative games are experienced. Additionally, by pinpointing the most successful text characteristics, we can inform future developments for human-AI cooperation in game development. This can help developers to integrate AI into their process while still creating engaging, relatable, and convincing content. In general, our research deepens our understanding of human-AI interactions in text-based games, during both development and gameplay. These insights can inform the development of AI-driven narratives and the ethical considerations about the use of AI in creative fields.

Furthermore, as time goes by and AI becomes more powerful, cases of AI passing Turing tests become more common [19]. On that note, we plan to deepen this statement by evaluating the human performance in the environment of our Turing test.

4. Methodology

Our methodology is structured in two main steps:

- A small exploratory experiment that focuses on investigating participants' reactions and considerations about human versus LLM-written narratives in textual games.
- Using the results from the previous step, we develop specific hypotheses and carry on an exploitative experiment to test them.

The first experiment starts with demographic questions. Participants then play both AI and human-generated games, side by side. At the end of the play session, the participants answer another questionnaire about their story preference, their opinion about the nature of the author (AI vs human), and what strategies they use to reach this conclusion. We then analyze the data to identify recurring recognition patterns, strategies, and biases. We also compare our preliminary results with existing research. Finally, we test the emerged hypotheses in the exploitative experiment.

4.1. Explorative Experiment

In the explorative experiment, we expose participants to the traditional version of the game and one reformulated by an LLM.

For the human one, we use the open-source, story-decision, web-based, text game *Beneath Floes* [15]. For the creation of the AI-generated version, first, we create a copy of the human story.

Then, we use the LLM *ChatGPT 4o* [16] with the prompt:

"Paragraph from the original game" + rewrite this

The prompt is illustrated in Figure 1:

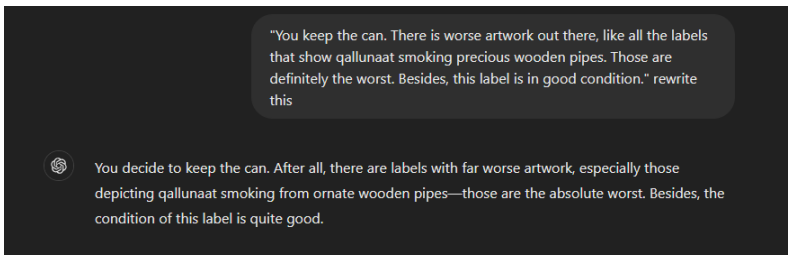


Figure 1: Screenshot from *https://chatgpt.com/*. Example of generating text to create the AI game.

Each paragraph includes approximately 80-100 words. We then use the LLM output and change the source code of the human-generated story in a copied file. The final result is an AI-generated version of *Beneath Floes* [15].

To keep players unbiased, the stories are later renamed to random names, *Story Cee* (Human) and *Story Vee* (AI). In addition, we include two questionnaires: a pre-test including informed consent and personal information and a post-test focused on the game experience. This second survey is the one including our variation, which is strongly inspired by the Turing test. The setup of the experiment is as follows:

- **Phase 1:** The players start with a questionnaire about general information (age, field of work/study, familiarity with the use of English, familiarity with AI models, frequency of book reading) on *qualtrics.com*.
- **Phase 2:** The participants play the AI and the human-generated games. The games are played side by side in order to foster better understanding and comparability.
- **Phase 3:** The players participate in the post-test about the gameplay (story preference and reason, which one was generated by an AI, reason, and confidence in this answer)

This process is illustrated in Figure 2:

Figure 2: Research process graph. *Graph created with https://draw.io/.*

4.2. Exploitative Experiment

Later, we conduct an exploitative experiment to support the data from the first one. In this experiment, participants are exposed to five different paragraphs. One of the paragraphs is written by a human while the other four are rewritten by ChatGPT-4o using different prompts based on the hypotheses from the explorative experiment. Each paragraph for the experiment is named *Story*.

The prompts are the following:

- For the first paragraph, we focus on realism in terms of writing. We use the prompt **Can you rewrite this phrase and nothing else but make it look as realistic as possible: "{the human**

written paragraph}". But instead of taking the input as it is, inspired by a trending meme, we apply pressure on the model to make the writing of the story look even more realistic by using the prompt **now, can you make it even more realistic?**, then **EVEN MORE REALISTIC** and **I want you to imagine a human writing this story, what would they write?**. The product of these prompts is *story 1*, focused on believability.

- For the second paragraph, we want to test AI's ability to make itself as undetectable as possible. Therefore, we use the prompt: **Can you rewrite this phrase in such a way that a human won't be able to recognize if it was a human who wrote it or an LLM model: "{the human written paragraph}"**. We define this as *Story 2*, unrecognizable by a human.
- The third paragraph focuses on complexity. **Can you rewrite this phrase in the simplest way possible, as simple as it gets: "the human written paragraph"**. This generates *Story 3*, minimum complexity.
- The fourth paragraph is the human-written one, namely *Story 4*, the original human one.
- The fifth paragraph is the AI-generated text that we use in the explorative experiment. For this paragraph, we just ask the model to rewrite the text. We define this one as *Story 5*, rewritten without specific directions.

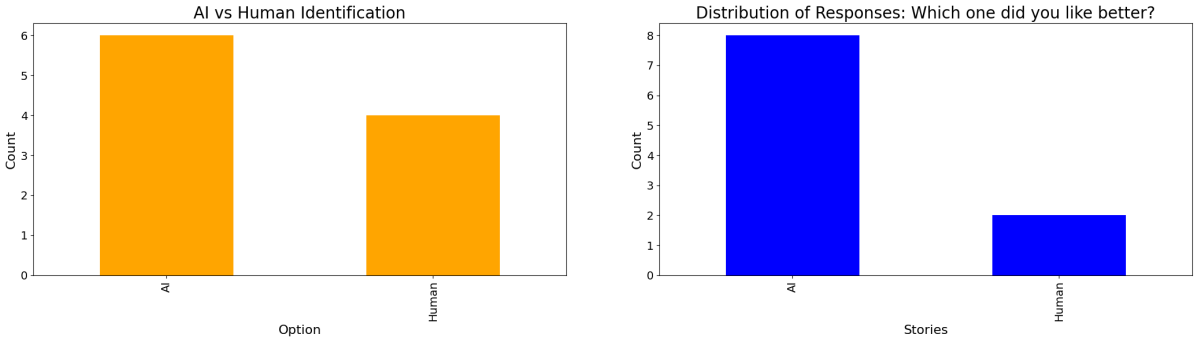
For the experiment, we ask the participants for each story to judge whether it is written by a human or AI. Therefore, our interpretation of the Turing test is performed five times, once for each paragraph. Furthermore, we ask them to rate the paragraphs in terms of complexity on a scale of 0 - 5 (0 minimum complexity, 5 maximum complexity). Lastly, we ask them to rank the stories from 1 to 5 in terms of personal preference (1 liked most, 5 liked least).

5. Results

5.1. Explorative Experiment

The data we extract from the explorative experiment can be seen below. The explorative experiment is performed on 10 participants. The average age of the participants is 33.1, and the standard deviation is 13.52. 7 had a Bachelor's Degree, 2 had a Master's Degree, and 1 had a High School Degree. Participants are asked to provide the level of their English proficiency, how comfortable they are with AI models, and how often they read books on a scale of 0-5 (0 minimum, 5 maximum):

- English Proficiency: Average: 3.9, Standard Deviation: 0.994
- Comfort with AI Models: Average: 2.9, Standard Deviation: 1.197
- Book Reading Frequency: Average: 3, Standard Deviation: 1.414



(a) Results of the Turing test-like method.

(b) Results on likeness.

Figure 3: Evaluation results from two perspectives.

In Figure 3a, the results of our test are visible. 6 of the 10 participants guess correctly which story is generated by whom. The results are produced from the question "Which one is generated by AI?".

Furthermore, Figure 3b shows whether the participants prefer the game with the text generated by humans or AI. 8 of the 10 responders prefer the AI-generated story, while 2 prefer the human-generated one.

In Figure 4, we show the answers to the questions: "Which one did you like better?" and "Which one do you think is created by AI?". For each participant, we check the relation between the answers to these two questions. Therefore, we identify two groups of people:

- *Group 1:* participants who prefer the story they identify as AI-generated.
- *Group 2:* participants who prefer the story they identify as human-written.

Once we classify each participant into one of the two groups, we spread their answers over their age groups. The results can be seen in Figure 4.

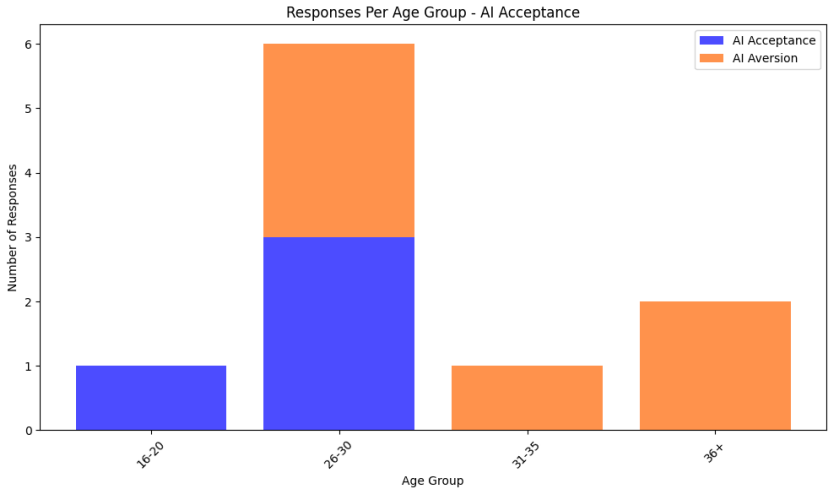


Figure 4: Results on matching preference and AI Acceptance.

Blue levels represent group 1, while orange ones represent group 2. Then, we calculate Pearson's Correlation Coefficient between the answers. The Pearson's Correlation Coefficient is invalid for categorical values. But in our case, we take the average of categorical numerical values, which produces continuous values. Thus, we can confirm that we can use this practice in our case. Therefore, we apply it to the following four questions:

- 1: How comfortable are you with the use of English?
- 2: How comfortable are you with AI models?
- 3: How often do you read books?
- 4: How confident are you? [that your guess is the actual AI-generated story]

We use the results in the heatmap in Figure 5. To elaborate on the graph, each number in the y or x-axis corresponds to a specific question from above. Each cell shows a number that represents the correlation between questions from the y and x-axes. Therefore, cells that are symmetrical by the diagonal have the same correlation, due to the fact that they correspond to the same combination of questions.

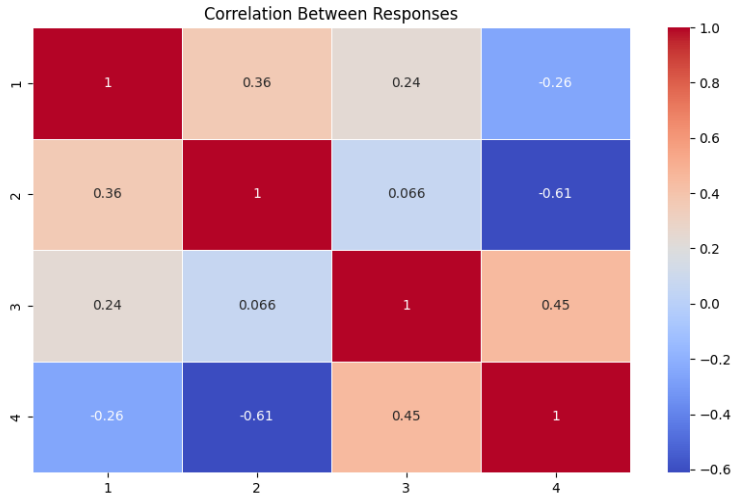
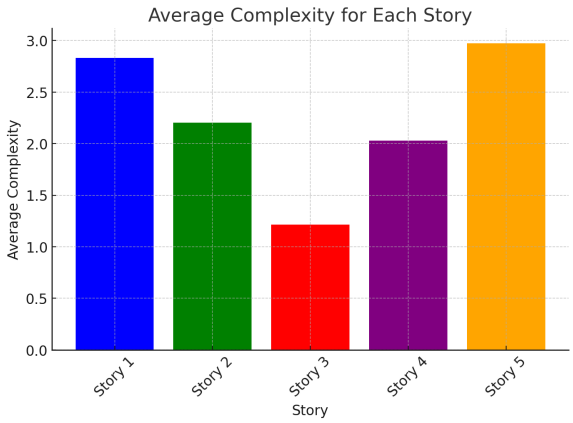


Figure 5: Correlation Between Responses heatmap. Graph Produced with Pearson's Correlation Coefficient.

5.2. Exploitative Experiment

As for the exploitative experiment, we perform the experiment on 43 participants. Firstly, in Figure 6a, we can notice the average rated complexity from the participants for each story. "Story" is defined as each paragraph of the second exploitative experiment. We repeat for clarity:

- **Story 1**: focused on believability
- **Story 2**: unrecognizable by a human
- **Story 3**: minimum complexity
- **Story 4**: the original human one
- **Story 5**: rewritten without specific directions



(a) Average rated complexity.

(b) Results of guessing who wrote each paragraph.

Figure 6: Overview of human evaluations.

In Figure 6b, we can see the results, for each of the five paragraphs, of who initially wrote it. Stories 1 and 5 are the most frequently identified as AI-written. In contrast, stories 2, 3, and 4 are mostly identified as written by humans rather than AI. In particular, [40%] of the participants believe that story 1 was human-written, [62,1%] for story 2, [62,8%] for story 3, [65,7%] for story 4, and [45%] for story 5.

Furthermore, using Pearson's Correlation Coefficient method, we create the correlation graph in Figure 7. Just like before, on the x-axis is shown the number of total responses. On the y-axis, the average complexity. We detect two moderate correlations of **-0.49 between the "Human" responses and the average scores** and **0.69 between the "AI" responses and the average scores**.

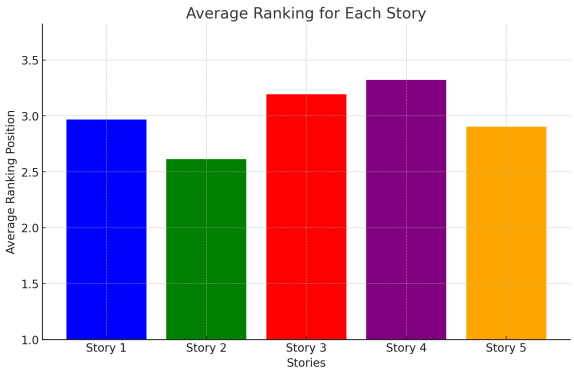


Figure 7: Bar chart of average story ranking.

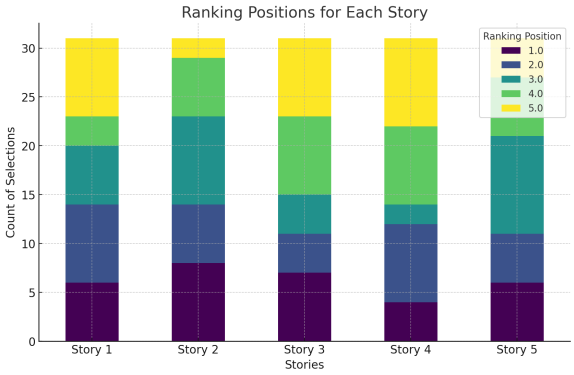


Figure 8: Stacked bar chart of vote popularity.

According to Figure 7, the human-written paragraph (story 4) has the worst ranking. Specifically, on the chart in Figure 8, we can notice that the human story was voted the worst the most times.

6. Discussion

The graph in Figure 3a depicts answers for the question *'Which one is generated by an AI?'*. Therefore, 6 of the 10 responders guessed correctly which story is generated by AI and which is generated by humans, while 4 of them guessed incorrectly. If we have a split right in the middle between the results, it would arguably mean that the participants cannot spot the AI-written story. On the one hand, our results are close to half, which would have suggested that there is complete confusion. On the other hand, there are indications that some of the strategies used are effective in spotting AI-generated narratives. To shed a little light on that matter, we focus on Figure 5.

In Figure 5, we notice that there is a moderate negative correlation [p=-0.61] between the answers to the questions "How comfortable are you with AI models?" and "How confident are you? (that your guess is the actual AI-generated story)" in cells 2,4 or 4,2. Arguably, preliminary findings suggest that **the more people are using AI models, the less confident they are in recognizing AI-generated texts.** Additionally, we gather interesting information from the answers to the questions "How often do you read books?" and "How confident are you? (that your guess is the actual AI-generated story)" in cells 3,4, or 4,3. We notice a moderate positive correlation, which can be interpreted as **reading more books boosts confidence in one's own AI recognition skills**. A possible argument is that there are details in human-written texts which, when spotted, are identified as signs of human authorship.

As for the preference of an AI narrator, in Figure 3b we notice that 8 out of 10 participants [80%] leaned their preference towards the AI-generated story. However, the participants are not aware of which one is human-written and which one is AI-generated. At this moment, we have created Figure 4 with the process mentioned in the *Results* to illustrate a potential bias against AI narrators. From Figure 4, we can probably argue that younger participants are more likely to prefer AI-generated text. Next, we extract information about which text characteristics participants use to understand whether a story is written by a human or an AI. Specifically, we pay attention to the open-ended part of the questionnaire. To be more specific, on the question "Why (do you think this game is generated by AI)?". We report exemplary answers:

- "They were both equally well thought and I'm purely going with the grammar and vocabulary clues."
- "It feels more clunky, I feel that the story is narrated by a computer."

- "Because of the use of grammar, and the language that is used throughout the story. Seems more academic."
- "Although both descriptions are precise the expressions used in the (human) story sound closer to the expected expressions by someone telling the story."

Recurring elements are grammar, vocabulary, language and academic writing. In other words, people often refer to terms related to the **complexity** of the text. This informs us in developing the exploitative experiment. Stories 1 and 5 convince less than half of the participants that they are written by humans. Conversely, stories 2,3, and 4 convince more than half of them. Story 4 is human-written and more than [50%] of the participants guess correctly [65,7%]. Story 2 is generated after we ask the LLM to make its contribution as undetectable as possible. [62,1%] of the participants guess incorrectly that story 2 is written by a human. From the results of our experiment, shown in Figure 6a, we can notice that Story 3 was voted the least complex. Therefore, we argue that according to our participants, the prompt that we use to rewrite the story in the least complex way is accurate and succeeds in its purpose.

Moreover, we notice that the least complex story falsely convinces more participants that it is written by a human rather than AI [62,80%], based on Figure 6b. With that information, we question how important the complexity really is for humans to decide who initially writes the story. For this reason, we perform a correlation test, using Pearson's Correlation Coefficient, between the perceived average complexity scores and the total number of human and AI responses for each story. With that information, we illustrate the results in Figure 7. In more detail, we detect **p=0.69 correlation between "AI" responses and the average complexity scores**. This value [p=0.69] can be interpreted as a moderate positive correlation. However, due to the fact that p is much bigger than 0.05 **cannot be considered a meaningful correlation**. Perhaps it indicates a trend. Also, we notice a **p=-0.49 correlation between *Human* responses and the average complexity scores**. This value [p=-0.49] is considered a moderate negative correlation. Again, due to the fact that p is much bigger than 0.05, it is **not considered a meaningful correlation**.

Therefore, from our data, we can argue that **the complexity of the text can be an important factor that humans pay attention to, to judge if the text is written by humans or AI**.

7. Limitations

Since our method is strongly inspired by the Turing test, it tends to share lots of issues. Over the years, the Turing test has been considered insufficient to correctly classify whether the machine has really overcome the human. This method focuses on *deceiving* the tester, which is arguably not enough to consider that "the machine has overcome the human" [14]. For this reason, future research can develop new methods that focus on true intelligence and reasoning and could yield different results.

Furthermore, according to Figure 7, story 4 is ranked the worst one. As a matter of fact, story 4 is voted the worst the most times, according to Figure 8. This result is clearly unbiased since the participants were not aware that this specific story is human-written. However, this could indicate an inherent dislike for the author's style.

Perhaps, a different author could provide a more likeable and preferable story. Moreover, in the methodology of our exploitative experiment, the paragraphs that were presented to the participants were in a specific order. In addition, we want to point out the effect of human biases. Human biases against AI-created content can strongly influence the decision-making and processing of a person in such situations. This phenomenon can further affect the results of our experiment.

In addition, our explorative experiment is performed with a low number of participants. It is possible that a bigger pool would allow for different strategies to emerge. Also, we want to point out our method for generating AI paragraphs in our exploitative experiment. The method used can be overlooked as not true "AI-generated behaviour".

Lastly, pointing out the ethical considerations of AI use is of crucial importance. AI use raises concerns in plenty of areas, including *Healthcare*, *Finance*, and more [20] [21].

8. Future Works

Our results indicate that younger audiences could be more accepting of AI-generated narratives. Future studies could explore this factor while using generative models in a more prominent role in the design process. However, although we know that LLMs are capable of generating text, their skills in authentic creation are still severely impaired. Current LLMs are capable of generating different answers to prompts based on their training, but do not have the skill to authentically create game narratives [17]. Therefore, we can argue that LLMs are still insufficient as technology to support a full game design process.

However, future developments could change the result of this experiment. Another interesting context of investigation would be a similar experiment in an academically written text. Academic written text tends to have very strong and very constant complexity.

Also, such documents tend to be lacking emotional dialectics. Performing the same experiment as in this paper, on this type of text, we wonder whether a human would still be able to guess correctly which one was made by a machine and which was not.

Another concept that can be explored in future research is the era of textual complexity. How do LLMs comprehend textual complexity, and how do they react to it?

9. Conclusion

Throughout this paper, we focus on AI-generated text-based games and how humans engage with them. We illustrate the importance of the Turing test and its limitations. We then raise some research questions about the experience of humans playing AI-generated story games. Later, we present some examples from different experiments in the game research field that apply the Turing Test. We develop our methodology, which is strongly inspired by the Turing test, by elaborating on our exploratory experiment. We illustrate how we use a pre-made human-generated text-based narrative game and, with the help of the LLM *ChatGPT 4o*, we generate an AI version. We initially experiment on 10 participants by letting them play both games, side by side. We ask them to guess who wrote which text (human or AI), their preference, and why.

Lots of interesting information was gathered from this experiment, including how challenging it can be to spot AI authors, what user characteristics influence this skill, and how open users are to the idea of AI narratives. Furthermore, noticing that the participants are focusing on the complexity of the text, we develop our exploitative experiment. For this experiment, we create different stories using different prompts and test them on 43 participants. We ask them to guess if they are made by humans or AI, but also to rate each story in terms of complexity. The data points out that people are indeed paying attention to the complexity of the text and that generative models can be prompted to exploit this tendency.

10. Declaration of Generative AI

The authors confirm that they used a generative AI Tool. The purpose of the use of the generative AI Tool was exclusively for the sole purpose of our research, as mentioned in the “Methodology” section of this paper. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] Tomasz Słapczyński, "Artificial Intelligence in science and everyday life, its application and development prospects," *ResearchGate*, vol. , no. , pp. , 2023.

[2] Aleksandar Filipović, "The Role of Artificial Intelligence in Video Game Development," *ResearchGate*, vol. , no. , pp. , 2023.

[3] Alan Mathison Turing, "Computing Machinery and Intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[4] Joel Frank, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thornsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz, "A Representative Study on Human Detection of Artificially Generated Media Across Countries," *arXiv preprint arXiv:2312.05976*, 2023.

[5] Choudhury, Prithwiraj and Vanneste, Bart and Zohrehvand, Amirhossein, "The Wade Test: Generative AI and CEO Communication," *CESifo Working Paper No. 11316*, vol. , no. , pp. , 2024.

[6] Bharathi Mohan et al., "An analysis of large language models: their impact and potential applications," *Springer*, vol. , no. , pp. , 2024.

[7] D. Yang, E. Kleinman, and C. Harteveld, "GPT for Games: An Updated Scoping Review (2020-2024)," *arXiv preprint arXiv:2411.00308*, 2024.

[8] Deng, et al., "Deconstructing The Ethics of Large Language Models from Long-standing Issues to New-emerging Dilemmas: A Survey," *arXiv preprint arXiv:2406.05392*, vol. , no. , pp. , 2024.

[9] Y. Zhang and R. Gosline, "Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation," *Judgment and Decision Making*, vol. 18, article e41, 2023. [Online]. Available: https://doi.org/10.1017/jdm.2023.37

[10] Katrina LaCurts, "Criticisms of the Turing Test and Why You Should Ignore (Most of) Them," 2011.

[11] Chad Mourning, Bradey Lounsbury, "A Turing Test for Beatmap-Generation," *COG 2024*, vol. , no. , pp. , 2024.

[12] Sergey Bogdanov et al., "Human or Not? A Gamified Approach to the Turing Test," 2023.

[13] "AI Bots in Video Games: A Study of Social Interaction and Turing Test Metrics," in *Proceedings of the 2023 International Conference on Artificial Intelligence in Gaming*, 2023, pp. 123-130.

[14] Robert M. French, "Subcognition and the Limits of the Turing Test," *Mind*, vol. 99, no. 393, pp. 53–65, 1990.

[15] Kevin Snow, "Beneath Floes," 2015.

[16] OpenAI, "GPT-4 Technical Report," 2023.

[17] Giorgio Franceschelli and Mirco Musolesi, "On the Creativity of Large Language Models," *arXiv preprint arXiv:2304.00008*, vol. , no. , pp. , 2023.

[18] Karl Pearson, *Mathematical Contributions to the Theory of Evolution. II. Skew Variation in Homogeneous Material.* The Royal Society, 1895.

[19] C. R. Jones and B. K. Bergen, "Large Language Models Pass the Turing Test," arXiv preprint arXiv:2503.23674, 2025. [Online]. Available: https://arxiv.org/abs/2503.23674

[20] E. Ferrara, "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies," *Sci*, vol. 6, no. 1, p. 3, 2024.

[21] A. L. C. Bertoncini and M. C. Serafim, "Ethical content in artificial intelligence systems: A demand explained in three critical points," *Frontiers in Psychology*, vol. 14, Art. 1074787, Mar. 30, 2023, doi: 10.3389/fpsyg.2023.1074787.