AI as Active Writer: Interaction strategies with machine-generated text in human-machine collaborative writing

ANONYMOUS AUTHOR(S)*

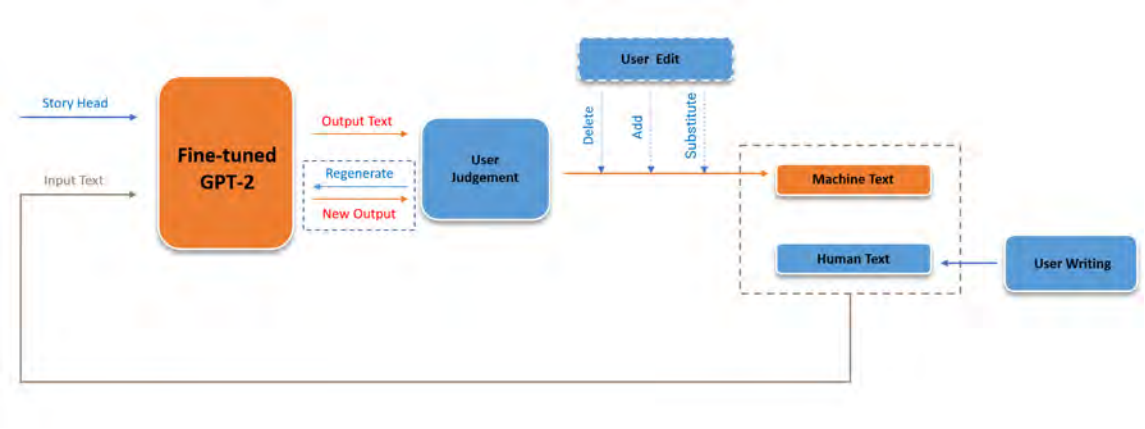

Fig. 1. Framework of Collaborative Writing System

Machine Learning (ML) has become an integral part of the creative process for human artists and writers, allowing practitioners to utilize a fuller source of information and be inspired by strategies and data previously seldom explored. To investigate the way human writers can collaborate with ML systems to generate creative fiction, we prototyped a web-based conditional text generation system that allows writers to shorten, edit, summarize, and regenerate text produced by AI. To investigate the dynamics of human-machine interaction in fiction co-writing, we used a "finish each-other's story" approach to allow humans and machine to take turns writing a collaborative fiction. We found that users took inspiration from unexpected text generated by the machine, that users expected reduced fluency and coherence in the machine text when allowed to edit the output, and that they perceived a mental model of the AI as an active writer in the collaborative process rather than simply as former AI writing assistant. This study provides framework and limitations for the way humans and machines can write creatively together in a future of increasing reliance on AI-augmented workflows.

CCS Concepts: • **Human-centered computing** → **Natural language interfaces**.

Additional Key Words and Phrases: Applications of intelligent user interfaces ; Collaborative interfaces ; User Modelling for Intelligent Interfaces ; Evaluations of intelligent user interfaces - Reproducibility

ACM Reference Format:

Anonymous Author(s). 2018. AI as Active Writer: Interaction strategies with machine-generated text in human-machine collaborative writing. 1, 1 (October 2018), 18 pages. https://doi.org/10.1145/1122445.1122456

Manuscript submitted to ACM

1 INTRODUCTION

The rapid development of machine learning has made it possible for artificial intelligence (AI) to collaborate with humans to generate creative content[9, 17, 20, 27, 34, 37]. In the foreseeable future, human-AI collaborative creative system based on machine learning will gradually enter people's creative artistic life, such as music composition[22, 34, 45], creative drawings[24, 37] and co-writing[7, 29]. These human-AI collaborative creation systems can assist experienced creators, such as inspiring creators and providing suggestions [6, 42]. At the same time, they can also bring a novel creative experience to users who have no or little creative experience in a short period of time, such as completing the drawing that the user has started or automatically filling in the user's unfinished sentence[7, 37]. In this article, we focus on the needs of users when they collaborate with AI for creative writing.

At present, a lot of work is focused on improving the algorithmic performance of natural language generation models, such as improving the logic of generated text[30, 40] or making the generated text closer to the natural language[15, 49]. However, little work focuses on exploring how users perceive the AI used for text generation and what interactive modes users want when working with AI for creative writing. Most designs consider collaborative creative writing systems with AI as the user's assistant by default, such as supplementing user unfinished sentences or providing users with suggestions for writing[6, 7, 29]. In this work, the premise of the machine as a writing assistant will be abolished, and the role of the machine in interaction will be studied. In addition, we explore what interactive capabilities users actually need when co-creative writing with AI, and how these capabilities affect the writing co-creation experience.

To ground our study, we developed a collaborative writing system with web-interface for human-machine co-writing. In this system, users and machine will basically take turns to write paragraphs for each other to continue with, and the system has two different modes, "Edit Mode" and "No Edit Mode", for users to interact with. Users are first required to give a beginning of a sci-fi story about human finding new homes. And a sci-fi theme fine-tuned GPT-2 model will give its story that is based on what users have written. Before continuing writing, users could regenerate, select machine generated texts. Additionally, in the "Edit Mode", users are allowed to edit the texts while in "No Edit Mode" they are not allowed. The machine would take every change that is made by users into account for its next generation. Users and machine will finally finish a 5 paragraph sci-fi story together, with 2 paragraphs generated by machine, and 3 paragraphs written by users.

By observing 9 users' writing process in two modes, surveying their opinions on machine generated texts and the co-writing process, interviewing their perceptions and feelings in the co-writing process, and analysing their written stories, we found that AI had good ability of generating unexpected twists such as adding new characters, new scenes, new events or changing the narrative atmosphere that made the users want to continue co-writing. But in recent stage, such unexpected twists should be refined by users in co-writing process to make it logic and coherent. We discovered that users with direct intend of their writing would have higher expectation on coherence of machine texts than users with implied intend. We also found that users tended to have higher fluency requirement if they could not edit machine texts. We revealed that most of users perceive the active writing machine in our system as idea generator but want to treat it as a human writing assistant or co-writer if the text quality could be improved.

In summary, we conclude our main contribution as follows:

1) We present the design and implementation of AI collaborative writing system that provide new interactive modes for users to co-write relatively long sci-fi stories with AI.
2) We find the patterns of texts in Human-AI collaboratively written stories: AI is a good unexpected twist provider but not a qualified writer.

3) We discover users with different writing intention as well as in different interactive modes (allow editing versus not allow editing) have different mental expectation on text coherence and fluency.
4) We describe users perceptions of machine's role in the co-writing process and discuss future possibility of writing machine.

Taken together, these findings guide the design of future Human-AI co-writing interfaces.

2 RELATED WORK

2.1 Text Generation Models

Recent breakthroughs in natural language generation (NLG) has established confidence in the field of human and machine language interaction. Due to the temporal nature of the language, a typical recurrent neural network language model (RNNLM)[41, 44] uses a recurrent neural network (RNN) to automatically regress the word or sentence input into a "hidden" vector and use this vector to predict the next word. Two variants of RNN, long short-term memory (LSTM)[18] and gated recurrent unit (GRU)[11], are proposed to solve that RNN cannot recognize the long-term dependence between words in the text. The sequence-to-sequence model (Seq2Seq) has also achieved great success in NLG[46], in which the encoder encodes text into vectors, and the decoder decodes these vectors and converts them into text. In addition, the model based on the Generative Adversarial Network (GAN) makes the generated text more realistic[15, 19, 36, 49]because GAN can generate synthetic data that is close to the real data.

The self-attention mechanism and transformer have opened up another path for natural language processing (NLP)[47]. The self-attention mechanism makes natural language processing out of the framework of RNNLM. The matrix calculation and embedding word vector in Transformer lay the foundation for large-scale parallel computing. Then the GPT model based on one-way transformer [38]and the Bert model based on two-way transformer[10] have been proposed one after another. Both have enormous training data and massive parameters and they are more suitable for complex NLG tasks than RNNLM. These two complex models can handle most tasks in NLP, such as NLG, text classification, question answering, and so on.

2.2 Strengths and Weaknesses of Machine Language Models

The flourishing development of Natural Language Understanding (NLU) provides a technical basis for human-computer collaborative writing. The language model of the machine understands the sentence relationship by predicting the next sentence of the binarization. It uses labeled data to train and classify different entity names to understand the meaning of words or phrases[35]. In fact, the machine learns the correlation between word vectors to calculate the importance between different words[10, 38, 47]. Therefore, language models understand data relationships rather than understand natural language. Although the way of understanding text by machine is different from humans' way[3], it injects new interest for human-machine collaborative writing. Compared with the difficulty of human innovation in story writing, machines cannot fully understand human writing intentions, so it is more likely to create unexpected plots and drive the development of the story. Due to the high-quality text training data, few errors occur in word spelling and grammar for machine[12, 33]. The machine based on the data relationship removes thinking time, which improves the efficiency in story creation. The training dataset of the machine is quite larger than the related knowledge in the human brain, so it has the potential to generate interesting story text[14, 32]. Therefore, human-computer cooperative writing will become the trend of future writing.

There are still many challenges in applying current machines to human-machine collaborative writing. The machine's understanding of the story text is based on the data relationship, so it cannot really understand the internal logical structure of the story text, which leads to the lack of robust logic in the machine-generated story text[5, 43]. In addition, various large-scale machine language models require huge training text data sets[10, 14, 38]. These training text data sets are a synthesis of languages, including various topics (science fiction, love, war, etc.) and various genres (fiction, news, dialogue, etc.). However, human-computer cooperative writing often has a fixed theme. This makes the machine-generated story text very likely to deviate from the subject[25]. The instability of long texts is also a problem faced by machine language models. Since the self-attention mechanism is mainly used in the current machine language model, in the process of generating long text, it is easy for the word vector to notice itself, thus falling into the looping state of the text[6, 7].

2.3 Human-AI Collaborative Creative Tools

The accelerated development of artificial intelligence has stimulated people's desire to explore the creation of cooperation between humans and AI. There are artificial intelligence in drawing creation[9, 24, 37], creative writing[7, 17, 29], dance[23] and other fields[13, 31]. For example, some use AI to complete sentences and provide suggestions[6] and others create music with AI[34]. In this series of work, AI acts as the user's collaborator. It can adjust its output according to the goals and actions proposed by the user and then make corresponding recommendations. In these systems, AI collects the user's input information as its output condition or predict the user's true intention based on the user's feedback. This system design maintains the consistency between AI and user intent and provides users with a comfortable experience at the same time.

Different AI tools shine in various fields of human-computer interaction. For example, Changhoon et al[37] designed DuetDraw based on Google's Sketch-RNN and PaintsChainer, which can complete the user's sketch, insert new objects on the vacant canvas, and fill the color according to user needs. Ryan and others[34] mainly used collaborative co-creation (COCOCO) designed based on Generative deep neural networks (DNNs) to realize the collaborative creation of music between humans and AI. Ammanabrolu and others[1] used the question-and-answer model in Albert to extract the basic information from the text ,combine them into a knowledge graph and then constructed an interactive novel world.

2.4 Human-AI Collaborative Creative Writing

At present, the mode of collaborative creative writing between users and AI can be mainly divided into text interactive games[2, 16, 21, 28] and writing assistants[4, 6, 7]. In a text interactive game, the user controls the character through natural language, and the AI agent recognizes the user's input, intelligently manipulates the character's actions in a text-described environment and feeds the results back to the user. AI agents can analyze the relationship between word embedding spaces to match objects and actions[16]. Optionally, the AI agent generates actions using a large number of predefined command patterns extracted through walkthroughs, tutorials, and decompiled games[28]. Designing skip-thought vectors[2] that include location information or effects with actions can also help AI agents understand the nouns entered by the user and return the corresponding actions.

AI writing assistant is also an important research field of human-AI creative writing. The AI writing assistant can correct users' spelling and grammatical errors, complete user's unfinished sentences or supplement full text paragraphs, and provide inspiration and suggestions for users' creative writing. Calderwood[4] and others designed an auxiliary writing interface that can complete sentences or paragraphs of text, and studied the writer's interactive feelings. There is

more interaction in the interface of Coenen et al[7], including multiple candidate continuation sentences, and rewriting or specifying selected text. However, most of the current research is only on auxiliary story writing or small sample story generation. Few work deeply explores the scenario of users and machines cooperating to write long stories. This is what we are exploring in this paper.

3 COLLABORATIVE WRITING SYSTEM

In order to study how human react with the sci-fi texts generated by machine, the collaborative writing system in this paper is aimed to be able to continue the story that is written by human, to be as easy to use as possible, and to be as easy to access as possible. As a result, we implemented fine-tuned GPT-2 model as the sci-fi text generator as shown in 1. First, the story head is written by humans and entered into the fine-tuned GPT-2. The user then judges the text generated by the machine and decides whether to regenerate it. After that, the text generated by the machine is modified by the user as the final machine text. The user follows the story development of machine text to write. Finally, both machine text and human text are used as input for the next machine generation.

3.1 Fine-tuned GPT-2

GPT-2 is a super-large-scale language model proposed by OpenAI in 2019[8, 39]. It has the ability to continue writing from short sentences into complete articles, and the model can find a suitable text style and add some details by itself. There are also some contextual connections in the generated text, and the structure is progressive. In addition, almost no grammatical and word errors appear in the text. Therefore, the GPT-2 model is suitable for our work.

GPT-2 has models of different sizes, including small (124M), medium (355M), large (774M), XL (1.5B). Since we need the system to recognize user input and continue to write the next paragraph of the story, GPT-2-small is not suitable for our work because the parameter is too small to meet our requirements for the logic, fluency and relevance of the generated text. In addition, GPT-2-large and GPT-2-XL were also abandoned by us because of the limitations of GPU and TPU. Finally, GPT-2-medium was applied in our system.

The original GPT-2 is trained on high-quality corpus from the Internet, which comes from the high-scoring external link pages appearing in the Reddit forum[39]. Therefore, the pre-trained GPT-2 is easy to generate concise and simple sentences, which are very close to news rather than stories or novels. In order to solve this problem, we fine-tuned GPT-2-medium in the field of science fiction and controlled the style of text generated by the model to be close to real science fiction. We chose the Sci-Fi Stories Text Corpus[42] collected by Robin Sloan as the dataset of fine-tuning GPT-2-medium. Most of the science fiction story data in it comes from *Pulp Magazine Archive*.

3.2 Operation System

3.2.1 Consider History. Most story generation models suffer from the fact that the generated text is not related to the original text. In 3.1, we used Sci-Fi Stories Text Corpus to fine-tune GPT-2-medium. This ensures that the text generated by the collaborative writing system is consistent with the overall theme of the science fiction and that the style of the paragraph is closer to the novel rather than the news. However, fine-tune can't connect the context of a specific story well and keep the themes consistency. In the collaborative writing system, we consider the story texts on consecutive time slices and package them as input in the next time slice . The long text input adequately contains the unique characteristics of different stories, which guides the language generation model in the collaborative writing system to generate text paragraphs that are more in line with the original story text.

3.2.2 Regeneration. Although fine-tuned GPT-2-medium has been able to generate realistic sci-fi story text, the model sometimes fails to generate readable text, such as repetition of text, wrong common sense of the world and unnaturally switching topics. These failed generated texts mislead users and greatly deteriorate the user's collaborative writing experience. If users cannot handle these failed texts, their passion for collaborative creation with AI will be frozen, because they have to face the torture of failed texts. Therefore, we provide users with the regenerate option to deal with failed generated text. When the user selects regeneration, the model re-selects a random seed and input it into the GPT-2-medium, in order to prevent the same generated text from appearing.

3.2.3 Edit. Although fine-tuning GPT-2-medium and setting Regeneration can ensure the readability of the machine-generated text to a certain extent, the unpredictable user aesthetics cause a huge gap in the evaluation of the text. Sometimes, parts of the machine-generated text or plot are preferred by the user, while the user does not satisfy the rest. We have introduced a edit function in the system to deal with these possible situations. We also want to figure out how users differently utilize the machine texts and perceive the machine's role with and without editing machine texts. In the editing box, users can freely add story lines, delete sentences they do not satisfy and replace words. The modification greatly improves the user's degree of freedom and we also discuss the impact of the edit on the collaborative writing of users and AI later.

3.2.4 Mode Choosing. The mode selection in the collaborative writing system gives users freedom to choose the interactive mode of collaboration with AI. In "Edit Mode", users can freely modify the text generated by the machine. In "No Edit Mode", users can only choose to continue story creation under the setting of machine-generated text, which may be suitable for users who challenge themselves. Different users prefer different interaction modes and show their creativity in their favorite modes. In addition, it is also convenient to explore how modes affect human-AI collaborative writing.

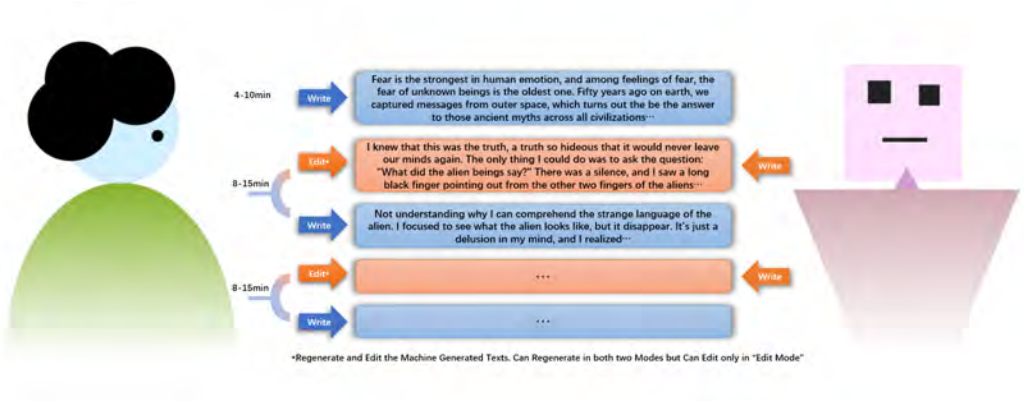

Fig. 2. Study Flow. Users and the machine take turns to co-write a short science fiction. Users first give a beginning of a sci-fi story of human finding new homes. Then after they get, regenerate, select, and edit the texts given by machine, they again write a paragraph to continue the story. They will repeat this to write paragraph 4 and finally end the story in paragraph 5.

3.2.5 "Edit Mode" and "No Edit Mode". As shown in 2, in both "Edit Mode" and "No Edit Mode", users have to input a beginning of a story to start generation. The beginning will usually be several sentences as paragraph 1. Then the GPT-2 model will take consideration of what they have written and provided a maximum of two hundred words to

continue the story as paragraph 2. Users are required to read these sentences and choose whether to regenerate the sentences or not. If they get the sentences they like, in "Edit Mode", they can edit these sentences as many as they desire, while in "No Edit Mode" they cannot. Then users are required to give another several sentences to continue the story as paragraph 3. After this, machine will take consideration of the history text and provide another a maximum of two hundred words to continue the story as paragraph 4. Again, users are required to read these sentences and choose to regenerate or not. In "Edit Mode", they can edit these sentences while in "No Edit Mode" they cannot. Then they are required to give another several sentences to end the story as paragraph 5.

3.3 Django Based Web Page Interfaces

The website is established based on Django framework and can be accessed through Internet. As shown in 3, two different interfaces are designed for two modes. "Edit Mode" and "No Edit Mode" both have "Submit" button for machine to get the human input, "Regenerate" button for machine to regenerate the sentences, and "End" button to end the story. "Edit Mode" has an additional "Edit" button for editing the text generated by machine. All history texts will be shown on the top side of the page, with human texts in black and machine texts in red. This would serve as a good reminder. The back button of the web browser allows users to go back to their last operation if they mistakenly operate the interface.

Fig. 3. Collaborative creative Writing system.**Left**:The initial interface includes writing prompts, mode selection, theme selection, input box and submit button.**Middle**:In the upper interface, the black font represents the text entered by the user, while the red font represents the text generated by the machine.After the machine generates the text, the user can choose to modify, regenerate, skip the modification to continue generating and end the interaction.**Right**:The user modifies the text generated by the machine.

4 USER STUDY

We conducted a user study to investigate the dynamics of human-machine interaction in fiction co-writing. To this end, we designed a collaborative writing framework that allows machine and user take turns to write a short science fiction story together. Additionally, we used two different interaction modes in the framework for users. While in one mode, users could edit what machine have written, in the other mode, users could not edit the texts generated by machine. We ask research questions (RQs) below in this study:

RQ1: What are the properties of the stories generated by a human-machine turn-taking system?
RQ2: What patterns of interactions are taken up by humans when they interact with machines in collaborative writing?
RQ3: How does the ability to select, edit, and cut out machine-generated text affect the human-machine co-writing process?

RQ4: How do humans perceive the role of the writing machine in the editable vs. noneditable modes interaction?

4.1 Measures

We surveyed particpants following playtest based on a 7-point Likert scale (1 referred to Strongly disagree, 7 referred to Strongly agree). Since users' past experience may have affections on their writing strategies with machine-generated text, we first asked about background and demographics: experience on science fiction writing: "*How much experience do you have in fiction writing?*" and experience on co-working with machine: "*How much experience do you have with machine learning technologies?*"

The quality of the machine generated texts may have great impact on story quality and users' co-writing strategies. Inspired by human evaluation metrics in machine learning field, we evaluated users' attitude to machine generated texts in following metrics that we most concerned. Fluency: Users rated "*How fluent is the machine generated text?*" Expressiveness: Users rated "*How expressive is the machine generated text?*" Unexpectedness: Users rated "*How unexpected was the text generated by the machine?*" Relevance: Users rated "*How relevant to the theme is the machine generated text?*"

To answer the research question 1 and 3, we evaluated metrics on users' feeling about their stories written collaboratively. Logical: Users rated "*How logically organized is the text you wrote collaboratively with the machine?*" Unexpected twists: Users rated "*How much unexpected twists were found in the collaboratively produced text you wrote with the machine?*" Coherence: Users rated "*How coherent is the story your wrote together with the machine?*" Expressiveness: Users rated "*How expressive do you feel the story you wrote collaboratively with the machine is?*" Usability: Users rated "*How easy is it to interact with the the collaborative writing system?*" Inspiration: Users rated "*How much was your own writing inspired by the text generated by the machine during collaboration?*" Participation: Users rated "*How much effect did your own writing have on the text generated by the machine?*" Attitude: "*To what extent would you like to continue writing with the machine?*"

For quantitative analyse, we implemented Wilcoxon signed-rank test to judge the significance of users perception difference between "Edit Mode" and "No Edit Mode". For qualitative findings, one experimenter conducted open coding analyse[26] to find the common characteristics in the users' story, thinking aloud and interview transcripts.

4.2 Method

We recruited nine users for our study, later referred to U1 to U9 in this paper. To ensure that they could complete the study tasks, we required that they had at least some English creative writing experiences, as shown in 4. Their writing experience on science fiction were vary from being novices (2 users), having a few experiences (3 users) to being fluent writers (4 users).

All users were notified what they would write before the test. We observed, surveyed, and interviewed their writing with the robot. Each user would finish two different sci-fi stories in two different interaction modes with machine. The survey and the interview would be conducted at the end of each mode for additional information about what we had observed during their writing. We demonstrated the whole procedure to all users on the interfaces that they would use to write. The maximum time of the test was one hour and a half, and the test was usually end within an hour.

4.3 Study Procedure

In our study, each user had to finish two different co-writing modes with machine. The two modes were labeled as "Mode 1" and "Mode 2" on the interface, referred to in this paper as "Edit Mode" and "No Edit Mode" respectively. The only difference between two modes is that users can edit the text generated by machine in "Edit Mode" but not in "No

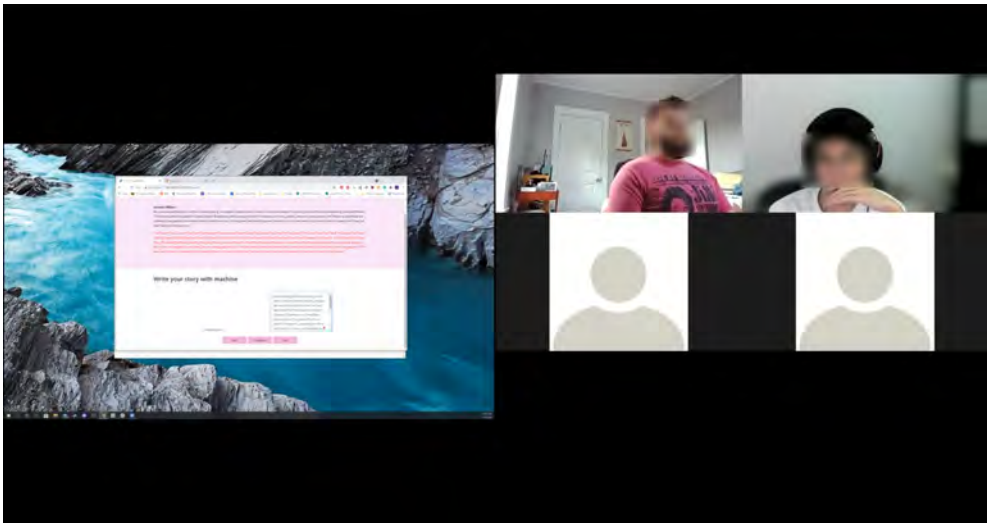

Fig. 4. Screenshot of the experiment. **Left:** The user share his screen. At this time the user is editing the text generated by the machine. **Right:** The user is in the upper left corner and the other three are experimenters.

Edit Mode". During the writing, users were asked to think aloud about their opinions on what machine had generated. After each mode, users were asked to fill out a 7-point scale Likert survey about what they had written. The two modes were counterbalanced as to which mode to begin first. Additionally, we would ask users for information that the survey might not cover.

The procedure went as follows:

1) Following a short demonstration of the "Edit Mode" user interface, they were asked to give their beginning of a sci-fi story about humans finding a new home in space. (Usually 4-10 minutes)
2) Users finished writing in "Edit Mode". (Usually 20-40 minutes)
3) They were asked to fill out a 7-point scale Likert survey on the machine generated text and the story written collaboratively in "Edit Mode". Then they were asked in semi-structured interview to answer some questions about the "Edit Mode" interaction. (Usually 5-10 minutes)
4) Following a short demonstration of the "No Edit Mode" user interface, they were asked to give another beginning of a sci-fi story about humans finding a new home in space. (Usually 4-10 minutes)
5) Users finished writing in "No Edit Mode". (Usually 20-40 minutes)
6) They were asked to fill out a 7-point scale Likert survey on the machine generated text and the story written collaboratively in "No Edit Mode". Then they were asked in semi-structured interview to answer some questions about the "No Edit Mode" interaction. (Usually 5-10 minutes)

5 QUANTITATIVE FINDINGS

The result of the two 7-point Likert scale surveys is shown in 5. The only significant differences in results' Wilcoxon signed-rank test[48] of two modes are the users' perception on fluency and expressivity of machine generated texts ($p=0.04858$ for fluency and $p=0.01788$ for expressivity). The result suggested that users had different mental expectation on machine texts in two mode of writing and thus selected texts in different qualities to continue their stories. In "Edit

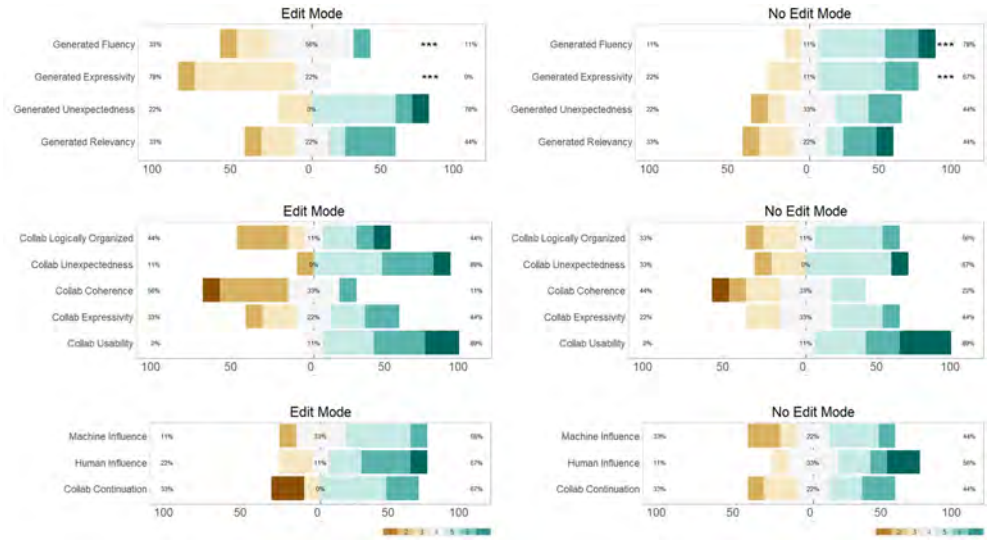

Fig. 5. Survey response data from Edit Mode and No Edit Mode testing. Perceived fluency and perceived expressivity of the machine generated text differ significantly between the groups. No significance is found in perceived interactive influence, nor properties of the collaboratively generated text.

Mode", users had lower expectation on machine texts since they could edit the texts. They would stop regenerating the paragraph if part of the paragraph could be used in their writing. In contrast, since editing was forbidden in "No Edit Mode", they had higher expectation on machine texts and tended to regenerate the paragraph until the whole paragraph could be used in their stories, and thus had more chance to encounter with high quality texts. There may also be effects on unexpectedness but there was not enough evidence to suggest significance (p=0.1675). Qualitative findings in 6.2 also provide evidence on the difference of their mental expectation on two modes. Indeed the end result of collaboratively generating stories in the two modes may not differ at all since assays of the collaborative texts generated in edit vs. no-edit yielded no significant difference in perception. Perceptions of being influenced by the machine text, or of humans themselves influencing the generation process appear to be the same in both modes, indicating that the editing options were perceived independently of the text generation process.

6 QUALITATIVE FINDINGS

In this section, we describe the properties of the co-written story, and user's strategies for co-writing with the active machine writer: what was their reaction towards the machine texts in two different modes, how machine texts affect their creative writing process, and how they perceive the partnership between them and the active machine writer in two different modes of writing.

6.1 Story Content

Based on the prompt, users tended to write about resource shortage on Earth in the context of space exploration: "*Humanity is stepping into to the age of space colonization right now, and the only problem stopping them is energy.*" (U8 "Edit Mode"), or "*Humankind finally found unity but at the price of their homeland. Nothing was left on the soil irrigated with bullets, shells, and flesh. Nothing to eat, nowhere to live.*" (U7 "Edit Mode"). Stories also tended to set limitations in

planet exploration or the space craft itself: "*The ship was designed to hold us for about 45, but already our resources are starting to be stretched thin.*" (U3 "Edit Mode"), or "*Their ship, once a proud and imposing vessel, now little more than a drifting wreck holed together with loose scraps and bolts.*" (U5 "No Edit Mode"). Aliens appeared in most of the stories, and they tended to be considered threats: "*It was as if a colony of ants were crawling up and down his arm just beneath the skin.*" (U2 "Edit Mode"), or be mysterious: "*Were them originally from another planet? I closed my eyes for a moment as I pondered this. I needed to know more, but when I opened my eyes, the creature was gone.*" (U1 "No Edit Mode"), or "*The men who had been with me all the time had changed. They were now men, and they looked like men.*" (U4 "Edit Mode"). Stories with threatening aliens always had negative descriptions about war: "*The war began soon when we reach Omega-2's outer surface. This alien race looks like huge disgusting worms, so we call them 'wormy'*" (U7 "Edit Mode"), or "*He could feel the bugs eating through his bones now. He weakly reached out and planted his hand on the self-destruct button.*" (U2 "Edit Mode").

Half users wrote first-person stories, using "*I*", "*We*" in stories. Half users wrote third-person stories, using self-designed characters as their protagonists like "*Commander Barone*" (U2 "Edit Mode"), "*passengers*" (U5 "No Edit Mode"), or "*Bibi*" (U9 "Edit Mode"). Users and machine all had a good grammatical person consistency: their description perspective would not change between paragraphs. Machine did this even better than users because one user (U6) suddenly made a change in the description perspective from third-person to first person without interpretation: the previous two paragraphs was about "*15 hand-picked young men were expected to go to Mars*" (U6 "No Edit Mode"), but the third paragraph started from "*But if we make...*" (U6 "No Edit Mode"), where "*we*" had unclear co-reference. This suggested that AI had compatible ability with human of focusing its writing on main character, which was the basic requirement of becoming a good co-writer.

New twists, including new characters, new scenes, new events were more frequently mentioned in selected machine texts (after regeneration, if any) than user texts. For example, in U4's "Edit Mode" story. U4 only mentioned "*We*" as new character, "*uncertain terrain*" as new scene, "*We are currently approaching a new solar system with a planet that seems inhabitable.*" as new event in paragraph 1. However, the machine wrote in both paragraph 2 and "*a man*", "*Icter*" as additional characters, "*winding corridor*", "*a tunnel*", "*a dimly lit room*" as new scenes, and "*walking down*", "*a man stood in front of me*", and "*should walk back and tell the others*" as new events. In the algorithm perspective, GPT-2 lack of long-time memory and logic ability made machine to frequently generate unexpected elements in the story. Furthermore, since the machine generated texts were selected by users, users also intended to remain such unexpectedness. Users regarded the unexpectedness as the core inspirations or reason of continuing their writing, and always took good use of the new elements, such as U4 wrote "*The others look at me inquisitively, wondering what was in the structure, and glad that I had made it out alright. I said 'there was a man.'*" in paragraph 3 after reading machine texts "*should walk back and tell the others*" in paragraph 2.

Although having many unexpected elements, user written texts and the selected machine texts were coherent with each other. Selected machine generated texts appear to use events, characters and scenes that were mentioned in user written paragraphs. For instance, since U2 had written "*However, one day an accident at the factory would force AB67 to do something extraordinary.*", machine continued the plot with "*And the result is this: a new robot, the first fully-autonomous, self-repairing, self-replenishing, fully-reactive, self-repairing robot.*", and mentioned the user-made word "*AB67*" in "*AB67 had been the first fully self-repairing robot.*" (U2 "No Edit Mode", user and machine wrote in paragraph 1 and 2). Similarly, users also did this in paragraphs that they wrote. For example, continuing the paragraph 2 in "Mode 2", U2 wrote "*AB67 didn't know it was the first fully-autonomous, self-repairing, self-replenishing, fully-reactive robot, and that's the way its makers wanted it.*" that utilized new machine given AB67's characteristics. Moreover, machine could even catch up with

implied info written by users, such as since U7 wrote a Cthulhu style beginning and wrote sentences "*The survivors were curling up in the corners, hands together above their heads praying for forgiveness, murmuring inaudible words.*" that described religious behaviors after "*In a blurring vision, I saw an indescribable, flesh-colored sphere with multiple eyes and tentacles. When I came to my mind, a slaughter has happened on the ship.*", machine continued with "*I knew that this was the truth, a truth so hideous that it would never leave our minds again.*" and wrote an alien saying "*Our religion is simple. We are a race of people who believe in a single God, and they are called the gods of the universe.*" (U7 "No Edit Mode", user and machine wrote in paragraph 1 and 2) that turned religious implicit into explicit dialogues.

More impressively, machine would sometimes suddenly change the positive atmosphere in previous paragraphs into negative atmosphere or reverse. For example, U2 wrote an optimistic beginning that "*As Commander Barone's shuttle hummed along, he couldn't help but feel a sense of optimism about humanity's future. He had successfully surveyed Planet T74 and was returning back to Space Station Endurance with a cargo hold full of samples of rocks, plants, and even some animal life.*", but machine suddenly turned into a negative description that "*He had been told that there were no known diseases or parasites on the planet.*" (U2 "Edit Mode", user and machine wrote in paragraph 1 and 2). Additionally, machine frequently tried to include two contradictory nuances in one paragraph. For example, machine wrote "*They had lost hope in all things that were human, in the hope of finding a better world, an era in which there would be freedom, happiness. But the crew of the Enterprise had never lost hope. They had found the way out of their nightmare. And it was to a world that they would return in years to come. But not in this world – this world of hope, freedom, and happiness. This was a world of death.*" (U5 "No Edit Mode", machine wrote in paragraph 4 ending) that gave reader a little hope and then destroyed it, or wrote "*This was maybe a good start, suddenly a big universal gravity like a storm "blow' us to a planet similar to Earth.*" that was positive, and "*They could tell us when we were coming from our star system to this home galaxy*" after aliens showed up that was also positive, but "*The most remarkable thing is that they didn't allow us back to Earth, or any other world.*" that was negative, then "*In the end, the people had to accept the fact that there had to be peace between the two super clusters, or they'd be destroyed.*" (U6 "Edit Mode", machine wrote in paragraph 2) that was positive.

6.2 Reaction to the Texts Generated by Machine

Users' reaction to the text generated by machine and their strategies of utilizing can be classified into two different groups by their expectation of the story they wrote: having direct intent about what they wanted to write (referred to as DI group), and having only implied intent about what they wanted to write (referred to as II group). Most users had concrete ideas about the story, saying things such as "*As in my mind. Earth is destroyed.*"(U2 in "Edit Mode"), or "*It changed my attempt of the story.*" (U6 "Edit Mode"), or *"It's kind of like I was planning to write: 'In the end, I found that main character was the murderer.'*"(U7 in "No Edit Mode"). By contrast, the users who had only implied intent would say like "I don't think about the ending "(U1 in "Edit Mode"), or "*I think I was more interested in working with, with the machine and like seeing what it did and then trying to build upon it.*"(U4 in "Edit Mode"), or "*I do not have a clear idea about what's going on.*"(U7 in "Edit Mode"). Users such as U7 may have different pre-expectant in different modes of co-writing. Users who had different predetermined expectation performed differently in the same interaction mode, and users who had the same expectation also performed differently in two interaction modes.

6.2.1 Reaction to Coherence of the Machine Texts. Users in II group had lower coherence expectant of the machine texts than users in DI group. And they all had more coherence requirements when they were doing "No Edit Mode".

Users in II group had requirements that texts should contain at least some new information that they could work on. For example, any new characters, events or locations could be good for them: "*I don't think I said anything about a name*

so I guess it named somebody, which is cool." (U4 in "Edit Mode"), or "*The machine generated that we have encountered the specific race of alien life. So that's a very good start for writing*" (U7 in "Edit Mode"), or "*It changed to the location all the sudden because it was outside and then now we're like inside. Because it's a winding corridor, I think.*" (U4 in "Edit Mode").

However, users in DI group were trying to found something that logically fitted to their story in the machine generated texts: expected subjects to have logical continuations such as "*I guess it depends first on what I wrote and then if I think it's a logical continuation.*" (U1 "No Edit Mode"), and refused illogical characters such as "*Machine starts spitting out more and more characters that were not mentioned in the scene which made it really wonky later on.*" (U5 "Edit Mode"). Additionally, they wanted the machine to continue the story that they expected, such as "*Well I expected the machine to basically take, you know, to see what I wrote and can expand upon it or relate to it in some way that's what I expected.*" (U5 "Edit Mode"), or "*More follow of what I had already made rather than add more twists*" (U3), or becoming frustrated when "*I was trying to give like a kind of optimistic adventure by, I think I use the word optimistic. And then the computer is like no, they're dying.*" (U2 "Edit Mode"). However, they would be even more excited if some unexpected items that logically fitted to their story appeared, like meeting with unexpected plot: "*And they took it even one step further with like, Okay, what if you peel off his skin.*" (U2 "No Edit Mode"), or with new characters: " *I think I could kind of use this because at first there wasn't an alien species in the original thing I wrote, so that was kind of a nice thing to add.*" (U3 "Edit Mode").

In "Edit Mode", both users in 2 groups would accept text that at least partial of it could use, like "*And if there are some sentences can use, you will definitely work, work, work on it.*"(U1 "Edit Mode"), or "*But I can work with these first three sentences.*"(U2 "Edit Mode"). However, in "No Edit Mode", users would expect the text to fully meet their requirements on coherence, like "*I think I definitely wanted something that flowed a bit better with a story, but with the first one, I was more okay with giving me something that perhaps added new ideas.*" (U3).

6.2.2 Reaction to Fluency of the Machine Texts. Fluency of machine texts was more important in "No Edit Mode" than "Edit Mode".

In "Edit Mode", for most of the users, partial readability would be their only requirement on fluency because "*if you're able to edit it and then it's less important because you can just fix it up a little bit of it.*" (U1). Nevertheless, the ratio of useful information the text conveyed might be key to judge fluency in "Edit Mode". Users always got rid of a machine generated text because "*It just said like basically the same thing over and over.*" (U3 "Edit Mode"), or *"The computer kind of get stuck in a bit of a loop.*" (U1 "Edit Mode"). Besides basic requirement, U5 had additional requirements on fluency that he expected machine to have some rephrase of some repeating words: "*The AI really likes to use the word ship, ship, ship.*" (U5 "Edit Mode", with frustrated tone); U6 had strict requirements on grammar accuracy: "*Some grammar mistakes, you know, will affect my understanding of the story.*" (U6 "Edit Mode"). However, U4 considered redundancy as a kind of metaphor and tried to make sense of it from the continued story: "*I feel like sometimes redundancy could be like a metaphor, or the meaning could be understood later.*" (U4 "Edit Mode").

In "No Edit Mode", fluency become as important as coherence for most users, such as "But if you can't (edit), then it's kind of more important that it is fluent." (U1). Similar to the difference mentioned in the part of coherence, users required the texts to be fully fluent if they could not edit it: "*There's stuff in all of them that like I kind of them, but it's hard to like make sense of the whole story. A lot of the times the computer gets stuck on word like 'diameter'.*" (U2 "No Edit Mode").

6.2.3 Reaction to Editing. All users agreed with that at least some basic edit should be allow to make the text to be more useful. The most common reason is along the lines of: "*This one is definitely harder because oftentimes there would be a good amount of it that would be useful and like I would want to keep writing off of. But then there's also be sections*

like a piece of sentences that were not great helpful." (U3 "No Edit Mode"). Even if some of the texts in "No Edit Mode" had high quality and met the strict requirements of users, most of users still felt editing was necessary, like *"I think some editing would be required because, You know, there's still some consistencies but not as glaring as that in first mode texts.*" (U5 "No Edit Mode").

Beside refining the text, U6 believed that by editing, "*I could add some thoughts about what I am thinking about the story and what I want the readers to know into the story.*" (U6 "Edit Mode"). However, U4 and U8 preferred editing not affecting the continued story. In another word, they preferred just refining fluency after the whole story have been generated, by saying: "*I wouldn't really change the story it comes up with but I would just change or delete a few sentences or something.*" (U4 "Edit Mode"), and "*I feel the story in 'No Edit Mode' have much more surprising twists. So the story is more interesting.*" (U8 "No Edit Mode").

Although all user agreed with that editing was essential, U6 preferred not to edit the text because editing was a burden: "*In the first mode ('Edit Mode'), I must understand the machine texts and then edit them. But in the second mode ('No Edit Mode'), I don't need to understand them and just choose one of my favorite and continue the story.*" (U3).

6.3 Redundancy and Unexpectedness

Due to limitations of the GPT-2 model with small corpus of pre-training, the AI frequently had repetitive descriptions on one single thing: "*For them the world was just a tiny speck, an insignificant speck of a speck, a speck, a speck of a speck in the vast emptiness of space.*" (U5 "Edit Mode", machine first generation on paragraph 2), or on one single sentence "*He gets a message, 'I am the only survivor,' he gets a message from a ship. 'I am the only survivor,' they reply with 'I am the only survivor.' ... so he gets the message again, 'I am the only survivor,' he gets a message from a ship. 'I am the only survivor, you are the only survivor.'*" (U8 "Edit Mode", machine first generation on paragraph 4). The repetitive descriptions was the main reason that led users to regenerate, because even if the paragraph was editable, it left little new information for users to work on. The situation might be much worse if the paragraph was not editable: the redundancy could even ruin the whole narrative. Users get visibly frustrated by these AI limitations if they could not get texts they want after a few iterations of regeneration. Besides, even the selected texts were usually not fluent enough for fiction before human refining, usually containing meaningless sentences that mislead user comprehension: "*We had been taken from the ship and taken from us.*" (U4 "No Edit Mode"), or had logic inconsistency: "*The boulder that I found was the largest I had ever seen. It was as large as a person's fist.*" (U1 "Edit Mode").

Although getting many redundant texts, users were amused when unexpected texts appeared, even if it presented some random events or characters that had no relationship to what users had written. For example, U2 laughed when he saw his story turned into a Christmas story, but regenerated it by saying "*This is not a Christmas story.*" (U2 "No Edit Mode"), or U8 "*think I'm quite happy to see some surprising twists. So the story will be more interesting.*" (U8). The unexpectedly redundant texts also amused users. For example, U4 laughed when encountering sentences like "*I was a human with a human face.*" in the story, and U5 laughed when saying "*That's a very odd sentence 'the man in the open suit it wasn't a woman', very weird.*" (U5 "Edit Mode").

Meeting with much unexpectedness, users just applied some of the unexpectedness that was easy to work with, or whether it drove the plot forward. In most situations, redundant texts were too hard to work with: "*I'm trying to like get some notes that fit a little bit more and gives the idea about how to drive the plot forward but it seems to like be redundant.*" (U5 "Edit Mode"). But there was one exception in U4's writing: "*I guess like the only way to make that sentence makes sense ('I was a human with a human face'), is if it wasn't redundant, the story could be that he didn't always have a human face.*" (U4 "Edit Mode"). Even when the text was not redundant, it could still be hard to work on when the plot was

being driven forward too quickly: "*I'm going to regenerate it because it focuses so much on death and yet I don't want it to be like at the start of the story.*" (U2 "Edit Mode"). However, this situation could be mitigated or even be useful if the machine wrote the ending: "*I feel like it wrote a decent ending on its own and didn't really want to add anything to it.*" (U5 "No Edit Mode", in delight tone).

6.4 Perception of Active Writing Machine's Role in Co-writing

Most of users perceive active writing machine as an idea generator. Its function might be generating hints that could be useful: "*The randomly generated text will somehow give me an idea on how to continue the story, but just, just like a hint. it is not very usable, but it can be used as a hint.*" (U7 "Edit Mode"), or generating ideas for writers when they get stuck: "*Maybe a machine could generate something that might give you an idea of where to take our story if you're feeling stuck*"(U1, "Edit Mode"), or even bridging gaps between writing milestones: "*But sometimes it's hard to connect them (milestones) together so with like random text generated, it might help a little bit to bridge those gaps.*" (U5). In addition to perceiving machine as an idea generator, U2 also thought that machine text could provide good writing exercise to open writers' mind: "*It was good for helping people who either want an exercise in like being agile with storytelling Or working on an improvisational type skill.*" (U2 "Edit Mode"). The reason was that "*it was good for forcing me to kind of be like nimble and rethink my concrete ideas.*" (U2 "Edit Mode").

One users (U4) perceived the machine's role differently as a co-writer. Interestingly, U4 never used the regenerate button to get different texts from machine in both "Edit Mode" and "No Edit Mode": "*I think I was more interested in working with the machine and seeing what it did and then trying to build upon it.*" (U4 "Edit Mode"). When being asked about the unexpectedness in machine texts, he answered "*if you were co writing with another person like another writers, they definitely bring a new perspective and new ideas.*" (U4), and he just accepted all unexpectedness and redundancy in the text produced by the machine.

However, all users agreed that at this current stage of development, machines may not be an adequate co-writer if it had to generate complete paragraphs by itself. They could accept part of the machine texts: "*I feel like it has its uses and it's definitely pretty interesting but I wouldn't want to rely on it 100 % on writing.*" (U1), and also "*adding certain sentences, but also not necessarily writing the entire thing on its own.*" (U3). U5 tried to treat machine as a co-writer, but he failed by saying, "*In this case it's AI and I are writing the story collaboratively, but I would not classified as a competent partner.*" (U5 "Edit Mode"). Even U4 mentioned that "*I think right now it seems not like a human writer but that it generates something helpful and then with a lot of editing you could use it.*" (U4 "Edit Mode"), and "*And just taking those pieces and rather than letting it have a whole paragraph by itself because it does a lot of redundant things and stuff.*" (U4 "Edit Mode").

7 DISCUSSION

7.1 AI's Writing Personality

AI's algorithm limitation made AI unpredictable to sometimes provide low quality texts full of words that could hardly make sense but sometimes high quality inspirations that might even move the plot forward beyond humans' expectation. Such personality made AI hard but fun to work with, and suggest that AI could potentially become a good writer after solving its redundancy problems.

As long as AI texts did not get stuck in a loop, AI showed a strong ability of providing unexpectedness. And it was even surprising to find that these unexpectedness conveyed literariness after human selection. AI's advantages in this regard appears to be on adding twists into the story, which could introduce excitement in the story. After selected

and refined by human writers, some unexpected but logical elements could make the story much more exciting than writers' previous intention. Furthermore, AI changing of atmosphere could serve as dramatic contradiction in the story that utilize positive foreshadowing to make negative scenes even more surprising. Additionally, AI had abilities to keep all unexpected twist under a theme, and had good grammatical consistency. These features give the AI a particular personality as an active writer in collaboration.

7.2 Affection of AI texts on Human-machine Co-writing process

Users did get a lot of inspiration from the unexpectedness provided by AI. However, the literary qualities of the machine text is perceived implicitly by users, who usually find coherent and fluent paragraphs coincidentally during generation. The cause of passive perception might be that the AI generating process was uncontrollable and nontransparent, and thus users could not have a reasonable expectation that could be satisfied, such as mentioned by U1 *"Maybe knowing that it's going to generate something that changes the way I approach writing so I don't think about the ending because I'm thinking about what what it might give me instead of what I think it should give me*" (U1 "Edit Mode"). This might also be the reason why they were excited when getting inspirations and frustrated when regenerating partially useful paragraphs.

Also, regenerating the bad quality texts from AI would lead users to lower their expectation and compromise on incoherent and tenuous text that conveyed merely inspirations: they may ignore some logic contradictions like "*There's a lot of weird contradictions in this paragraph, but at least it's like, in keeping of the theme of the story.*" (U2 "No Edit Mode" after 4 times of regeneration), or may even stop the regeneration " *Let's see the one or two more and then pick whatever it generates*" (U5 "Edit Mode").

7.3 Possible Future Role and Future Interaction Modes of Writing Machine

No matter how users perceived the partnership between machine and them, better quality of machine texts would allow them to focus more on the idea generated rather than the functionality of the system. Since users could regenerate machine texts in both modes, most of them perceived the machine as an active idea generator used to find inspiration. They preferred "Edit Mode" more since they could pick what they loved regardless of the fluency of the texts. Thus, it is important to give the right of edit to users. Both editing machine texts and users' texts should be allowed at any time in the writing process. Furthermore, since users may have different expectation on various machine text characteristics, more controllable variables can be added into the system for users to control the machine texts for incorporation into the story, reducing redundancy. For example, using conditional text generation, the system could allow users to control the numbers of new characters and scenes, or the atmosphere of the plot, or some weights that could help AI to focus on some important part in user written plots.

In addition to treating machine as active idea generator, U5 wish the machine to be a human writing assistant. In this case, machine should be able to accept both previous paragraphs and following paragraphs as inputs to connect milestones in the story for him. Also, some personal inputs should be added to make the generated machine texts more controllable.

Besides, U2 and U4 regarded machine as an active co-writer or a writing exerciser. In this case, the texts were expected to be uncontrollable. But they also had basic requirements on quality of machine texts: less redundant and few grammar mistakes. However, sometimes redundancy could be recognized as metaphors. The definition of good quality of texts could be vague in this mode of interaction. More research should be conducted to develop a good co-writing or writing exercise machine.

8 CONCLUSION

We found that AI had potential to become a decant co-writer under human supervision in co-writing process and revealed its recent limitation of being a redundancy generator. We discovered that different mental expectation of users could affect their strategies and their perception of machine's role in co-writing process. Future systems should both improve the output text quality of AI generated texts, keep advantages of AI generating unexpected twists and take users mental expectation into consideration to achieve better interaction effectiveness. Taken together, this work advances the frontier of Human-AI co-writing interfaces, enabling users to utilize the advantages and avoid the disadvantages of AI writing.

REFERENCES

[1] Prithviraj Ammanabrolu, Wesley Cheung, Dan Tu, William Broniec, and Mark Riedl. 2020. Bringing stories alive: Generating interactive fiction worlds. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 16. 3–9.

[2] Timothy Atkinson, Hendrik Baier, Tara Copplestone, Sam Devlin, and Jerry Swan. 2019. The text-based adventure AI competition. *IEEE Transactions on Games* 11, 3 (2019), 260–266.

[3] Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5185–5198.

[4] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study.. In *HAI-GEN+ user2agent@ IUI*.

[5] Elizabeth Clark, Yangfeng Ji, and Noah A Smith. 2018. Neural text generation in stories using entity representations as context. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2250–2260.

[6] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*. 329–340.

[7] Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a Human-AI Collaborative Editor for Story Writing. *arXiv preprint arXiv:2107.07430* (2021).

[8] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164* (2019).

[9] Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, and Brian Magerko. 2016. Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 196–207.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Rahul Dey and Fathi M Salem. 2017. Gate-variants of gated recurrent unit (GRU) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*. IEEE, 1597–1600.

[12] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197* (2019).

[13] Alex Elton-Pym. 2020. Principles for AI Co-Creative Game Design Assistants. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 16. 335–336.

[14] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* (2018).

[15] William Fedus, Ian Goodfellow, and Andrew M Dai. 2018. Maskgan: better text generation via filling in the_. *arXiv preprint arXiv:1801.07736* (2018).

[16] Nancy Fulda, Daniel Ricks, Ben Murdoch, and David Wingate. 2017. What can you do with a rock? affordance extraction via word embeddings. *arXiv preprint arXiv:1703.03429* (2017).

[17] Katy Ilonka Gero and Lydia B Chilton. 2019. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[18] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems* 28, 10 (2016), 2222–2232.

[19] Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018. Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[20] Matthew Guzdial, Nicholas Liao, Jonathan Chen, Shao-Yu Chen, Shukan Shah, Vishwa Shah, Joshua Reno, Gillian Smith, and Mark O Riedl. 2019. Friend, collaborator, student, manager: How design of an ai-driven game level editor affects creators. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[21] Matthew Hausknecht, Ricky Loynd, Greg Yang, Adith Swaminathan, and Jason D Williams. 2019. Nail: A general interactive fiction agent. *arXiv preprint arXiv:1902.04259* (2019).

[22] Cheng-Zhi Anna Huang, Hendrik Vincent Koops, Ed Newton-Rex, Monica Dinculescu, and Carrie J Cai. 2020. AI song contest: Human-AI co-creation in songwriting. *arXiv preprint arXiv:2010.05388* (2020).

[23] Mikhail Jacob and Brian Magerko. 2015. Interaction-based Authoring for Scalable Co-creative Agents.. In *ICCC*. 236–243.

[24] Pegah Karimi, Jeba Rezwana, Safat Siddiqui, Mary Lou Maher, and Nasrin Dehbozorgi. 2020. Creative sketching partner: an analysis of human-AI co-creativity. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 221–230.

[25] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* (2019).

[26] Shahedul Huq Khandkar. 2009. Open coding. *University of Calgary* 23 (2009), 2009.

[27] Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. 2019. May AI? Design ideation with cooperative contextual bandits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[28] Bartosz Kostka, Jaroslaw Kwiecieli, Jakub Kowalski, and Pawel Rychlikowski. 2017. Text-based adventures of the golovin AI agent. In *2017 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 181–188.

[29] Max Kreminski, Melanie Dickinson, Michael Mateas, and Noah Wardrip-Fruin. 2020. Why Are We Like This?: The AI architecture of a co-creative storytelling game. In *International Conference on the Foundations of Digital Games*. 1–4.

[30] Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2021. ProoFVer: Natural Logic Theorem Proving for Fact Verification. *arXiv preprint arXiv:2108.11357* (2021).

[31] Yuyu Lin, Jiahao Guo, Yang Chen, Cheng Yao, and Fangtian Ying. 2020. It is your turn: collaborative ideation with a co-creative robot through sketch. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[32] Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[33] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504* (2019).

[34] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. 2020. Novice-AI music co-creation via AI-steering tools for deep generative models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[35] Alireza Mansouri, Lilly Suriani Affendey, and Ali Mamat. 2008. Named entity recognition approaches. *International Journal of Computer Science and Network Security* 8, 2 (2008), 339–344.

[36] Weili Nie, Nina Narodytska, and Ankit Patel. 2018. Relgan: Relational generative adversarial networks for text generation. In *International conference on learning representations*.

[37] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[38] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[39] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[40] Chang Shu, Yusen Zhang, Xiangyu Dong, Peng Shi, Tao Yu, and Rui Zhang. 2021. Logic-Consistency Text Generation from Semantic Parses. *arXiv preprint arXiv:2108.00577* (2021).

[41] Mittul Singh, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2020. Subword RNNLM approximations for out-of-vocabulary keyword search. *arXiv preprint arXiv:2005.13827* (2020).

[42] Robin slogan. 2016. Writing with the machine. *https: //www.robinsloan.com/notes/writing-with-the-machine/* (2016).

[43] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203* (2019).

[44] Minguang Song, Yunxin Zhao, Shaojun Wang, and Mei Han. 2021. Word similarity based label smoothing in RNNLM training for asr. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 280–285.

[45] Minhyang Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[46] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[48] RF Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* (2007), 1–3.

[49] Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. 2017. Adversarial feature matching for text generation. In *International Conference on Machine Learning*. PMLR, 4006–4015.