Fostering Creative Story Writing Through Personalized Question-Asking Agents

Niki Hasrati
University of Waterloo
Waterloo, Ontario, Canada
nhasrati@uwaterloo.ca

Ehsan Jahangirzadeh Soure
University of Waterloo
Waterloo, Ontario, Canada
ejahangi@uwaterloo.ca

Brian D. Zimmerman
University of Waterloo
Waterloo, Ontario, Canada
bdzimmer@uwaterloo.ca

ABSTRACT

In creative writing classes, asking students specific questions about their writing is a commonly employed technique used by teachers. In addition, this method is also theorized to help reduce writer's block. However, personalized questions on one's work is currently unavailable to aspiring writers without access to a teacher. In this work, we design an accessible agent that asks students personalized questions to encourage them to reflect on specific aspects of their written stories. Through a survey of the literature on creative writing, we identified two broad categories of questions, which we refer to as story- and reader-based questions. We conducted a pilot experiment with 12 participants with no professional creative writing experience, who received either story-based questions, reader-based questions, or no questions while writing a story. Preliminary results show that our questions led participants to further develop their stories, and that story-based questions were preferred over reader-based questions. In addition, contrary to our hypothesis, participants in the agent conditions experienced more writer's block compared to participants in the baseline condition.

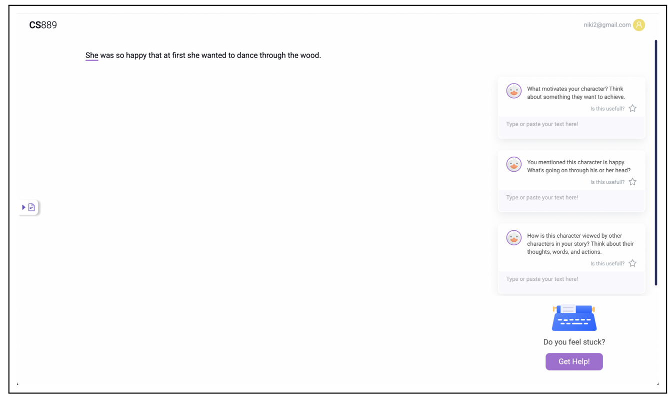

Figure 1: Prompter, a personalized story writing assistant.

1 INTRODUCTION

Storytelling is a fundamental aspect of human nature, forming the basis of our social interactions [25, 27, 49]. However, beyond simply being a communication tool, storytelling also provides many benefits for both children and adults. Telling stories constitutes an important part of children's cognitive and linguistic development [46] and is used as a pedagogical technique for increasing their literacy [29]. For adults, creating personal and emotional narratives can help in the recovery of patients with post-traumatic stress disorder [5] and has even been shown to improve the physical and mental well-being of otherwise healthy individuals [31, 39]. One of the most popular ways to engage in storytelling is through creative writing, which offers additional benefits since both creativity and writing have been identified as essential skills for personal and professional success in the twenty-first century [4, 44].

Despite the importance of creative writing, there is a lack of effective resources to help individuals foster their skills. The most popular and successful learning environments for creative writing are in-person and workshop-based [10, 22, 41]; however, such programs are rarely part of the teaching curriculum past elementary school [3] or outside of specialized post-secondary courses [41]. In addition, evidence suggests that a crucial aspect of these programs' success is the quality of the personalized feedback offered by skilled instructors [22], which has been shown to suffer as class sizes increase [13]. As a result, our aim is to make the development of creative story writing skills accessible to a wider audience by building a tool that provides personalized feedback supported by pedagogical theories of creative writing and storytelling.

Previous tools have focused on generating automatic continuations of the users' writings [9, 23, 34], asking users to enact a story before writing it [8, 40, 42, 50], implicitly influencing users through time constraints [2] and subliminal priming [17], or enabling social collaboration on writing projects [15, 20]. However, none of these tools meet all three of our criteria: accessible, personalized, and grounded in pedagogical theories.

To make the development of creative writing skills more accessible, we present *Prompter* (displayed in Figure 1), an agent that asks personalized questions based on the user's writing. These questions are designed to prompt users to reflect on their own text, and are modelled based on existing methods for teaching creative writing, as well as insights gleaned from the writing process of successful authors. Despite evidence to support that systematic question-asking is an effective method for reducing writer's block [28] and supporting creative writing skills [6, 7, 26, 35, 37], to the best of our knowledge, this is the first work exploring the efficacy of personalized question-asking agents in this domain.

we conducted a pilot experiment with 12 computer science graduate students with no extensive prior experience with creative writing. Participants received either story-based questions, reader-based questions, or no questions while they wrote a story. The story- and reader-based question categories were identified from a survey of the literature on creative writing. Preliminary results show that the questions led participants to further develop their stories and that story-based questions were preferred over reader-based questions . However, contrary to our hypothesis, participants in the agent conditions experienced more writer's block compared to the baseline condition. Our work contributes an accessible and personalized story writing support tool as well as insights into pedagogical methods for fostering story writing skills and reducing writer's block.

2 RELATED WORKS

2.1 Pedagogical Theories

2.1.1 *Creative Writing.*

Due to the persistence of romantic myths that creative writing is an innate talent, there is little research on how best to teach creative writing [41]. As a result, to learn more about relevant pedagogical techniques, we turned to two sources: descriptions of creative writing teaching methods used in graduate-level degrees [6, 26, 35], and accounts of the writing process of professional authors [7, 37]. Through a review of these sources, we identified two high-level categories for questions designed to encourage novice writers to reflect on their work: story-based questions, which prompt writers to focus on different elements of their story (e.g. how a story character will react to an event in the story), and reader-based questions, which prompt writers to anticipate the reactions of their readers (e.g. how the reader expects the story character to react to this situation). This categorization may seem counter-intuitive at first glance; however, it is backed by psychological theories of narrative comprehension. Studies in cognitive and neuropsychology have shown that when reading a narrative, the reader constructs a mental representation of the story's characters and goals, and it is this mental simulation that leads to emotional responses while reading [25]. These findings shed light on why some authors have an intuitive understanding of the importance of imagining the reader's response while they are writing [37]. However, we are not aware of any other works comparing the story-based and reader-based mental models during the writing process.

The prompts used by writing instructors cover a variety of different story-elements, which range from characters and settings to more abstract concepts such as plot, tone, and pacing. Given research on the neuropsychology of narrative comprehension, which suggests that characters' goals and intentions are what give a narrative its significance [25], we decided to limit the scope of our study to character-based prompts. Based on the works of Rogers [35] and Card [7] we identified the following categories to prompt users to think about: the actions that the character takes, the character's motivations, the character's past, the dialogue and speech of the character, the character's thoughts, other characters' views of the character, and the character's physical description. For each of these categories Prompter will either ask users to think about the character directly or how the character would be perceived by someone else reading the text.

2.1.2 *Feedback.*

Past work has shown, in a variety of different teaching domains, that receiving personal feedback is an important aspect of students' learning [18]. Such feedback is also conducive to fostering creative writing skills specifically [22]. However, personalized creative writing feedback is difficult to access past elementary school [3] or outside of specialized post-secondary courses [41], and even then, effective teaching suffers when class sizes become too large [13]. We built Prompter to provide *formative feedback*, which supports and scaffolds learners in their trajectories to achieve their desired goals [45]. In this way, Prompter emulates a social context in which a learner is scaffolded by a "more capable peer" [45].

2.1.3 *Writer's Block.*

Writer's block is an affliction encountered by novice writers in which they experience a sense of helplessness and an inability to continue writing [36]. Oliver [28] proposes three methods for addressing writer's block: helping students to develop their ideas through question asking, prompting students to write freely, and encouraging students by providing them with feedback that highlights their strengths as writers. Prompter is designed to reduce the occurrence of writer's block based on Oliver [28]'s first two suggestions: by systematically asking users questions about their writing as well as reducing distractions and providing a brainstorming area to allow free-writing. In this study, we do not consider Oliver's third suggestion, since automatic evaluation of story quality is as of yet an open question.

2.2 Writing Support Tools

Outside of the story writing domain, adaptive tools for improving argumentative writing skills have shown potential for scaling personalized feedback to wider audiences [47, 48]. These tools provide adaptive support to teach users how to make more compelling arguments. This adaptive support is catered to what the writer has written and fosters a context-sensitive feedback loop. The findings surrounding these tools are encouraging since users found them useful and were motivated to continue honing their writing skills afterward. As a result, writing support tools have shown promise in providing accessible, personalized feedback in writing tasks other than story writing, and Prompter is predicated on the success of support tools in other writing sub-domains.

2.3 Creativity Support Tools in HCI

Inspired by the notion that computers have the potential to enhance human creativity [12], and that the design of such tools is a grand challenge for HCI [38], Creativity Support Tools (CSTs) have become an active research initiative [30]. These tools aim to lower the threshold of entry into creative domains and/or provide inspiration to stimulate the creative process [2]. However, compared to tools supporting the creation of music or the visual arts, creative writing support tools have received less attention [2].

2.3.1 *Creative Story Writing Support Tools.*

Some creative writing support tools focus on helping with specific aspects of creative writing, such as generating metaphors [16] or word choices [14]; however, we are specifically interested in tools that aid in the writing of narratives as opposed to stylistic features of the language. One category of such tools use neural language models to suggest automatically generated continuations of the text [9, 23, 34]; however, these methods are aimed more as a test of the current state-of-the-art abilities of neural language models as opposed to contributions to pedagogical theories of writing. Other categories of story writing support tools involve those that use

Table 1: Examples of Prompter's Questions

Characterization Techniques	Story-based Prompts	Reader-based Prompts
Physical Description	What's something unique about this character's appearance?	Can you show the reader something unique about this character's appearance?
Actions	How do these characters interact when they're feeling upset?	Can you reveal, through how they interact, that these characters are feeling upset?
Motivations	Why is this character feeling angry?	How can you show why this character's feeling angry?

time constraints [2], subliminal priming [17], or acting the stories out first before writing them [8, 40, 42, 50]. Although these works contribute to our understanding of writing pedagogy, none of them provide users with individualized feedback based on their writing. The last category of story writing support tools we identified enable social collaboration between different writers [15, 20]; however, there is evidence to suggest that beginning writers are reluctant to share their writing [22], making such tools inaccessible to novice writers.

2.3.2 *Design of Creative Writing Support Tools.*

Clark et al. [9] have identified three design considerations for creative writing support tools: interaction structure, interaction initiation, and interaction intrusiveness. Below, we will define each criterion and discuss our design choices.

Interaction structure can be *iterative*, where the piece of writing is refined in multiple steps, or *additive*, where the writing is added to but not revised. Clark et al. [9] viewed story writing as a sequential task and hence chose an additive interaction structure that locked in the user's text and disallowed editing after each computer suggestion was incorporated. However, it seems to be unanimously agreed that revision is a crucial step in any writing process [1, 10]. This is reflected in Hemingway's famous saying that "[t]he only kind of writing is rewriting" [19]. As a result, we designed Prompter to have an iterative interaction structure, allowing users to reassess our suggestions and their writing whenever they wish.

Clark et al. [9] define interaction initiation as either automatically initiated or user-initiated. They chose the suggestions of their writing assistant to be automatically initiated. However, two-thirds of their participants mentioned that "they did not find the suggestions helpful once they had a clear direction for the story." This may explain why 8 of 9 of their participants disregarded the system's suggestions. Furthermore, Gonçalves et al. [17] showed that users prefer receiving writing prompts at the beginning of the activity or when blocked. Based on these findings, we designed Prompter's suggestions to be user-initiated.

Finally, interaction intrusiveness refers to whether the writing assistant's suggestions can be ignored or not. Clark et al. [9] designed their system's suggestions to be highly intrusive; however, to promote distraction-free pre-writing, which is an important stage in the writing process [28], our prompts are only displayed when the user needs help or inspiration. Furthermore, to allow searching for ideas without necessarily including them in the final piece, Prompter gives users a place to brainstorm, making the suggestions unobtrusive.

3 SYSTEM DESIGN

3.1 Design Rationale

We designed Prompter to help improve storytelling and reduce writer's block. This was done by considering the following factors in the design of Prompter: the mode of delivery for the prompts, the type of prompt, and the intended writing process. The rest of this section will explore our design decisions for each point.

3.1.1 *Prompt Delivery.*

First, we designed the prompts' delivery to be user-initiated. That is, the user will have the option to press a button when they need inspiration or help. This decision was based on findings from prior work showing that users of creative writing support tools prefer receiving suggestions only when they are blocked or cannot think of new ideas to write about [9, 17]. In addition, by suggesting potentially distracting and unwelcome prompts, we did not want to interfere with *free-writing*–writing freely without pausing to evaluate or judge one's work—which is an important part of the writing process and may also help to reduce writer's block [28].

3.1.2 *Prompt Type.*

Our second consideration in the design of Prompter was the type of prompt. Through a review of the literature on creative writing methods and the writing process of professional authors, we identified two high-level categories for questions designed to encourage novice writers to reflect on their work: 1) *story-based* questions, which prompt writers to focus on different story elements (e.g. how a story character will react to an event in the story), and 2) *reader-based* questions, which prompt writers to anticipate the reactions of their readers (e.g. how the reader expects the story character to react to this situation) [6, 7, 25, 26, 35, 37]. In writing workshops, beginning writers are asked both story-based and reader-based questions about different elements of their story, for example, characters, setting, and plot [6, 26, 35]. In this work, we focus on character-based prompts, but future work should investigate the effects of prompts on other story elements. Based on this categorization, we have created and evaluated two character-focused agents, one story-based and the other reader-based. Our prompts are based on several techniques for characterization that we identified from the literature on story writing pedagogy [7, 35]. Examples of these techniques include the exploration of a character's physical description, dialogue, actions, and habits. More specifically, in Prompter, each sentence will be automatically analyzed to detect features of interest about characters, which the agent will prompt the user to reflect on. For example, if the user mentions that "Sarah was excited," the story-based agent could ask,

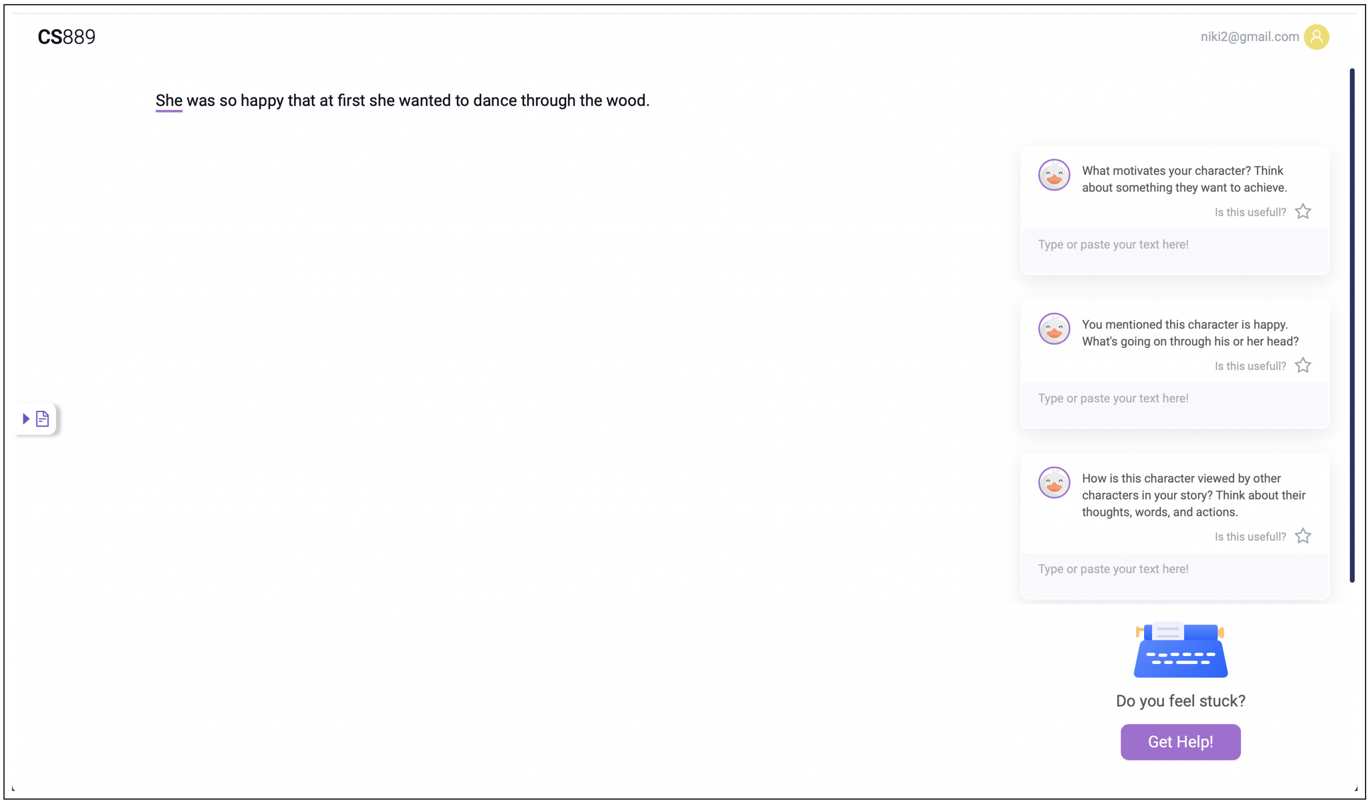

Figure 2: Prompter, a personalized story writing assistant.

"Why is Sarah excited?" while the reader-based agent could ask, "How could you show the reader that Sarah is excited?" Not only do these questions have the potential to improve the quality of written stories, systematic questioning may also help reduce writer's block [28]. Examples of Prompter's questions can be seen in Table 1.

3.1.3 *Writing Process.*

Lastly, our design also takes into consideration the type of writing process we wish users to experience. Writing instructors and professional authors emphasize the importance of brainstorming and pre-writing in the creative writing process [26, 28, 35], as a result, Prompter provides users with a space to brainstorm possible answers to our personalized question. Furthermore, by allowing users to include as much detail as they want into the brainstorming area instead of directly into the story, this feature is also intended to reduce over-description, which is viewed as a negative quality in creative writing [10, 35, 37].

The rest of this section will describe how our creative writing assistant tool, Prompter, was built to provide personalized feedback on stories written by users. To make our system more accessible, we designed Prompter as an interactive web application as shown in Figure 2. Users interact with Prompter by typing or pasting a snippet of their story into the main writing area. When in need of help or inspiration, feedback can be requested by pressing a button, at which point the input text is sent to our personalized prompting algorithm in the back-end. This algorithm will analyze the input text and provide an appropriate prompt to display to the user. The following sections describe the design of the user interface and feedback algorithm in more detail.

3.2 User Interface

Prompter consists of three main components: a text editor for writing stories; prompt cards for displaying the personalized questions; and note-taking boxes for brainstorming ideas, keeping notes while writing, and responding to the prompts. The user interface can be seen in Figure 2.

3.2.1 *Writing Area.*

To keep users focused on the writing task as well as to reduce distractions, *Prompter's* writing area is a simple text editor that does not include extraneous functionalities for formatting the text. In addition, the writing area is designed to be the largest element on the screen to promote writing.

3.2.2 *Prompt Cards.*

Users can request a prompt from our system by pressing the help button. These prompts are phrased as questions about a specific part of the user's writing and are designed to simulate ideas and encourage reflection. When the help button is pressed, the user's text will be sent to our personalized prompting algorithm for analysis. *PromptScript* will return a single question about the user's text as well as the location in the text that the question is referring to. The question returned by *PromptScript* is then displayed in the UI in a *prompt card*, which consists of three components: the text of the question, a star button for soliciting feedback from users, and a text field for brainstorming ideas before incorporating them to the final text. In addition, using the location returned by *PromptScript*, the section of the text that the question is referring to is highlighted. Clicking on this highlighted section brings to focus the prompt card tied to it and vice versa.

3.2.3 Notes.

In addition to the above-mentioned note boxes, which are tied to the prompts, Prompter also provides a general note panel. This note panel gives users a place to keep track of their ideas, since it may be difficult to start writing directly without brainstorming first.

3.3 Personalized Feedback Algorithm

At a high level, our feedback algorithm analyzes the user's text to extract meaningful information and uses this information to populate a set of hand-crafted question templates. For our pilot study, we focus on extracting information about story characters. Specifically for each animate entity in the text, we extract all references to that entity, the start and end index of each entity reference, whether the entity is singular or plural, whether the entity is named, and the emotion tied to that entity, if any. To get all valid questions relating to the user's text, we populate a relevant subset of the question-templates using the above-mentioned extracted information as well as the prompt type (story-based or reader-based). Based on the questions that have already been shown to this user, we select one question out of the valid subset to return to the front-end. All natural language processing and information extraction is implemented using the Stanford CoreNLP library [24], which we interact with through a Python implementation, Stanza [32]. Our process for extracting character information is described in more detail below.

3.3.1 Character Detection.

For character detection, we first use Stanford CoreNLP's coreference resolution annotator to obtain all mentions to a specific entity in the text, including proper nouns and pronouns referring to it. For example, the text "There was a wooden chair. It didn't look comfortable" would yield two entity mentions, "The chair" and "It," tied to a single entity ID. We use coreference resolution as opposed to named entity recognition, a more common method for identifying entities, since we wish to extract unnamed entities (e.g. "the boy") as well as named entities (e.g. "Sam"). A limitation of our work is that Stanford CoreNLP's coreference resolution is not entirely accurate, leading to inconsistencies in entity recognition. Furthermore, we use CoreNLP's 'animate' tag to identify whether an entity is animate, filtering out dates, locations, and inanimate objects. However, CoreNLP's animacy detection has some flaws, leading to some inanimate objects being mistaken for characters.

3.3.2 Emotion Detection.

For emotion detection, we first use the Stanford CoreNLP's OpenIE annotators to extract, for each sentence, possible subject-object tuples. If the subject of one of these tuples corresponded to a character reference as described above, and the object of the tuple matched a hand-crafted list of emotions, the emotion was added to that character reference. This method allows us to correctly extract the emotion even when the subject and object are separated by multiple words (e.g. "I was so very frightened."). However, there are limitations as we are only able to extract emotions that exist in our list.

3.3.3 Prompt Generation.

Given the user's text, our feedback algorithm chooses a subset of hand-crafted question-templates to present to the user. For each character reference extracted from the text, the possible questions are selected based on three criteria: whether the character reference is tied to an emotion, whether the character is singular or plural, and whether the prompt type is story-based or reader-based. If the character reference is not tied to an emotion, our algorithm has access to 15 question-template categories which focus on general questions pertaining to the characters. When passing these questions to the front-end, they are tied to the first reference of that character. If the character reference is tied to an emotion, 6 question-template categories are available. These questions are tied to the specific location of the character reference which is tied to the emotion. For each of these question-template categories, there are 4 options depending on the prompt type and the singularity status of the character.

3.3.4 Prompt Selection.

For each user, we store all questions that have been previously shown to them, along with the text of the character reference that the question refers to. Out of all possible prompts generated based on the user's text, we randomly select one to display; however, we never repeat the same question text, even across two different characters, unless all unique question texts have been exhausted. When all of the possible prompts, based on the input text, have been displayed to the user, our algorithm returns an empty response.

4 STUDY DESIGN

4.1 Research Questions

Since we designed Prompter based on pedagogical techniques for improving story writing and reducing writer's block, we are interested in assessing Prompter's efficacy in these domains. In addition, we are also interested in comparing the effects of different prompt types. Specifically, we ask the following research questions (RQ):

- **RQ1:** *How effective is Prompter, compared to a baseline, in helping users write better stories?*
- **RQ2:** *How effective is Prompter, compared to a baseline, in reducing the occurrence and frequency of writer's block?*
- **RQ3:** *How is the usefulness of Prompter affected by the type of prompt used?*

4.2 Procedure

4.2.1 Methodological Overview.

To answer our research questions, we ran a controlled between-subjects experiment with 12 participants and three conditions: 1) story-based agent, 2) reader-based agent, and 3) baseline. First, all participants were asked to complete a pre-study survey. This survey gathered information on the personal characteristics of the participants and their previous experience with creative writing. Controlling for these measures, the 12 participants were then randomly assigned to either the control group or one of the agent groups. All three groups were given 10 minutes to write a story based on a pre-selected starting prompt. While writing, participants in the two agent conditions (story- and reader-based), had access to all of Prompter's features and could request help to receive either story- or reader-based prompts depending on their condition. For the baseline, participants also wrote the story using our system, but they did not have access to the help button or any features other than the text editor. All components of our study were carried out asynchronously to reduce the effects of observer's bias on students' writings.

4.2.2 Participant Selection.

Participants for the pilot study were selected from students in an HCI graduate course at a North American university. All 13 students in the course were invited to participate, 12 of which completed the study. Participants were compensated with partial course credit.

4.2.3 Demographics and Participant Assignment.

The 12 participants completed a pre-study survey which asked about personal demographics and past experience with creative writing. Out of the 12 participants, 58.3% identified as female and the remaining as male. English was the first language of 41.7% of the participants. All participants stated that they were at least comfortable writing in English, with half of all participants being very comfortable. When asked about their comfort level in writing stories, 33.3% stated that they were not comfortable, 50% stated that they were comfortable, and 16.7% stated that they were very comfortable. In terms of story writing experience, 58.3% of participants had written some stories in the past, 41.7% had never written any stories, with none of the participants being professionally published or serious hobbyists. Controlling for gender, comfort with writing English, comfort with writing stories, and extent of story writing experience, we randomly assigned 4 participants to each of the three conditions.

4.2.4 Instructions.

Participants were provided with a link to the system, anonymized credentials to log in, instructions on how to use the system, and overview of what to expect during the study. Since the study was carried out asynchronously, participants were also informed that the could pause the study at any time by pressing a 'pause' button. Those in the agent conditions were provided with additional instructions on how to request, star, and comment on prompts.

4.2.5 Writing Task.

After logging into the system, participants were presented with a note asking them to write a story about the following premise: *"Please write a story about a child who is having a bizarre day at school."* Participants were given 10 minutes to complete their story and had the option to pause the study timer by pressing the 'pause' button. Participants received reminders at 5, 3, and 1 minute left to the end of the study. Those in the agent conditions could request prompts, star prompts, and use the brainstorming note panel at their discretion.

4.2.6 Survey.

After the study timer had reached 10 minutes, participants were presented with a link to a post-study survey, consisting of a combination of open-ended and close-ended questions. The close-ended questions were 7-point Likert scales, with 1 *Strongly Disagree* and 7 representing *Strongly Agree*. All participants were asked about their experience with writer's block, their satisfaction with the quality of their story, and their feedback on the system. Participants in the agent groups were also asked about the usefulness of the prompts they received.

4.3 Data Collection

4.3.1 Session Data.

During the writing task, we collected log data from users' interactions with our system. This included the content and time of all keystrokes in both the main writing area and the note taking sections. We also stored the number of prompt requests for participants in the agent conditions. For each prompt, the question text, the location of the question, whether the prompt was starred, and all notes tied to the prompt were stored. The number study pauses was also recorded.

4.3.2 Self-Perceived Story Quality.

After completing the writing task, participants in all conditions were asked four close-ended survey questions on a 7-point Likert scale. Example questions included "I think my story is creative" and "The characters in my story feel like real people." We also asked an open-ended question asking participants to elaborate on which aspects of their story could be improved.

4.3.3 Self-Perceived Writer's Block.

Subjective measures of writer's block were collected through six questions on our post-study survey, which users in all three conditions are asked to respond to. All questions were followed a 7-point Likert Scale. Example questions included "I frequently ran out of ideas while writing" and "It was hard for me to continue writing after getting stuck." We also asked participants to write a free-form description of a time when they were stuck while writing and why.

4.3.4 Perceived Prompt Usefulness.

To evaluate how different prompts affect the writing process, we have designed two version of Prompter, one with story-based prompts and one with reader-based prompts. To compare the effects of the two agents, one measure we collect is participants' perception of the usefulness of the prompts while writing. During the post-study survey, participants in the two agent conditions are asked seven 7-point Likert scale questions such as "The prompts helped me improve my story" or "The prompts were distracting." Participants were also asked to describe a time when the prompts helped them and a time when the prompts were not helpful.

4.3.5 System Usability and User Feedback.

To measure general usability, we asked the following 7-point Likert scale questions: "I felt the system was easy to use," "I enjoyed writing the story," and "If given the chance, I would continue using this system in the future." In addition, participants were asked if they had any comments on how the system could be improved.

5 EVALUATION AND RESULTS

To evaluate our hypothesis that receiving individualized writing prompts during the writing process will help foster creative writing skills, we aim to answer the following research questions (RQ):

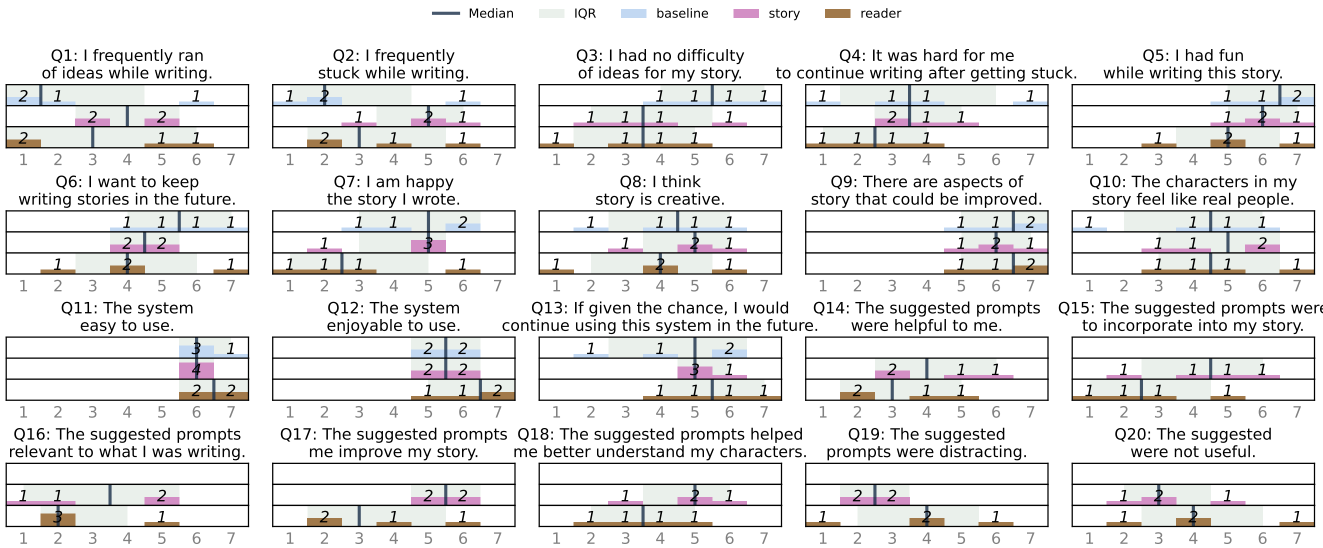

Figure 3: Frequency Distributions of Survey Responses

- **RQ1:** *How effective is Prompter, compared to a baseline, in helping users write better stories?*
- **RQ2:** *How effective is Prompter, compared to a baseline, in reducing the occurrence and frequency of writer's block?*
- **RQ3:** *How is the usefulness of Prompter affected by the type of prompt used?*

The first research question will be answered by comparing automated and self-perceived measures of writing quality across participants in the three conditions. To evaluate the second research question, we compare the perceived experience of writer's block as well as behavioural log data between the three conditions. The third research question will be addressed by comparing the survey responses and system interactions of participants in the two agent groups. To assess whether the differences between the groups are statistically significant, we perform a two-sided Man Whitney U test for the close-ended survey responses, and a double-sided t-test for the log metrics. All tests of statistical significance were carried out using the SciPy library [43]. The open-ended survey responses were coded by one researcher using thematic analysis. We also performed two-sided Mann-Whitney U tests to verify that, in terms of responses to our pre-study survey questions, there were no statistically significant differences among participants in the three conditions . For all questions and between all conditions, we received p-values larger than 0.25, implying no significant differences based on our measures of user demographics.

5.1 Writing Quality

5.1.1 *Quantitative Survey Responses.*

Participants rated the quality of their story on a 7-point Likert scale (1: strongly disagree, 7: strongly agree). Figure 3 (Q7-Q10) shows the frequency distributions of responses compared across the three conditions. First, we compare the results of Story Prompter to the baseline. As seen in Table 2, from the medians of the responses between the two conditions we observed that Story Prompter participants tended to judge the quality of their stories more highly compared to baseline participants. However, using a two-sided Mann-Whitney U test, we compared the results from each of the questions and did not find that theses differences were statistically significant (all four p-values were greater than 0.65). It is nonetheless promising that for questions 7, 9, and 10, the median values for Story Prompter are better than the neutral value of 4. Second, we compare the the results of the second agent condition, Reader Prompter, to that of the baseline. We observed that Reader Prompter participants were less satisfied with the quality of their stories compared to baseline participants. Again, none of these differences were statistically significant (all four p-values were greater than 0.23). .

Table 2: Self-Perceived Writing Quality

Question	Story	Baseline	Reader
Q7: happiness with story	**Med = 5.0** (IQR = 1.5)	**Med = 5.0** (IQR = 2.5)	Med = 2.5 (IQR = 3.0)
Q8: creativity	**Med = 5.0** (IQR = 1.5)	Med = 4.5 (IQR = 2.5)	Med = 4.0 (IQR = 2.5)
Q9: room to improve	Med = 6.0 (IQR = 1.0)	Med = 6.5 (IQR = 1.5)	Med = 6.5 (IQR = 1.5)
Q10: characters	**Med = 5.0** (IQR = 2.5)	Med = 4.5 (IQR = 3.0)	Med = 4.5 (IQR = 2.5)

5.1.2 *Automated Metrics.*

To the best of our knowledge, there exists no consensus on automated metrics for evaluating the quality of a story. As a result, we selected a few interesting metrics and evaluated them to assess story quality. The descriptive statistics for these metrics can be seen in Table 3.

Table 3: Automated Measures of Story Quality

Metric	Story	Baseline	Reader
Distinct Tokens	Mean = 93.25 (SD = 31.73)	**Mean = 145.25** (SD = 37.30)	Mean = 94.99 (SD = 21.56)
μ-TF-IDF	Mean = 0.0106 (SD = 0.0016)	Mean = 0.0125 (SD = 0.0007)	Mean = 0.0116 (SD = 0.007)

Distinct Tokens: Distinct tokens refers to the number of unique words in each story. More distinct tokens in a story could represent a more robust vocabulary for describing characters and other story elements. No statistically significant relationships were found between the groups but baseline participants generally had a more extensive vocabulary in their stories.

TF-IDF: *Term-Frequency Inverse Document Frequency* (TF-IDF) is a common metric for scoring the informational value of a particular token to a document in a corpus. Term frequency (TF) represents the number of occurences of a token t in a document. Inverse document frequency (IDF) is the logarithmic inverse of the number of documents containing the token t. The TF-IDF score of t in a document is the product of the term frequency and the inverse document frequency for that token. In evaluating participant stories, we take the average TF-IDF across all words for each story. In stories where unique vocabulary is frequently repeated, TF-IDF scores will be higher and indicate a more unique story.

Although stories written by participants in the baseline group had much higher μ-TF-IDF, the difference between groups was not statistically significant. We performed a multivariate analysis of variance on story length and μ-TF-IDF but found nothing conclusive within 5% ($p \approx .07$)

5.1.3 *Qualitative Survey Responses.*

We asked participants to write about aspects of their story they thought could be improved. Analyzing this data, we found three recurrent themes for areas of improvement: 1) writing style, 2) story elements, and 3) more time.

Writing style: 4 of our 12 participants expressed that their writing style could be improved, both in terms of vocabulary and grammar. Examples of these concerns include finding more appropriate words to express ideas and including more diverse sentence patterns.

Story elements: Half of the participants believed that story-specific elements of their writing could be improved. The frequency of these concerns was the same across all three conditions. Examples of story-specific story elements include the setting, the characters, the storyline, the story events, and the ending. Interestingly, only participants in the baseline condition expressed concerns with their story setting or environment. This may be because all of Prompter's prompts focused on characters, so for participants in the two agent conditions story setting may have been less salient.

More time: 6 of our 12 participants stated that having more time to write would have improved the quality of their story. 4 of these participants mentioned that the lack of time meant that their stories were not detailed enough, and 3 believed that the time constraint hindered their ability to finish the story. This is an important finding, and suggests that future studies exploring creative writing support tools should ensure that adequate time is given to complete the task.

5.2 Writer's Block

5.2.1 *Quantitative Survey Responses.*

Participants rated their experience of writer's block on a 7-point Likert scale (1: strongly disagree, 7: strongly agree). Figure 3 (Q1-Q4) shows the frequency distributions of responses compared across the three conditions. First, we compare the results of Story Prompter to the baseline. As seen in Table 4, from the medians of the responses between the two conditions we observed that Story Prompter participants' self-reported experience of writer's block was higher than that of the baseline participants. However, using a two-sided Mann-Whitney U test to compare the results from each of the questions, we did not find that the differences between the two conditions were statistically significant (all four p-values were greater than 0.23). Second, we compare the the results of the second agent condition, Reader Prompter, to that the baseline. The results in this case are more mixed, with Reader Prompter participants reporting more writer's block as measured by Q1-3, but less as measured by Q4. Again, none of these differences were statistically significant (all four p-values were greater than 0.11).

Table 4: Self-Perceived Writer's Block

Question	Story	Baseline	Reader
Q1: lack of ideas	Med = 4.0 (IQR = 2.0)	**Med = 1.5** (IQR = 3.0)	Med = 3.0 (IQR = 4.5)
Q2: getting stuck	Med = 5 (IQR = 1.5)	**Med = 2.0** (IQR = 2.5)	Med = 3.0 (IQR = 3.0)
Q3: no difficulty thinking of ideas	Med = 3.5 (IQR = 2.5)	**Med = 5.5** (IQR = 2.0)	Med = 3.5 (IQR = 2.5)
Q4: difficulty continuing after getting stuck	Med = 3.5 (IQR = 1.5)	Med = 3.5 (IQR = 3.5)	**Med = 2.5** (IQR = 2.0)

5.2.2 *Automated Metrics.*

We also looked at several automated proxy measures of writer's block, which we extracted from log data of participants' interactions with the system. Table 5 shows descriptive statistics for these measures compared across our three study conditions. The rest of this section will discuss how we defined each measure as well as report our findings.

Story length: We measured each participant's final story word count as a proxy measure for how much writer's block they were experiencing. We observed that participants in the baseline condition wrote stories that were, on average, approximately twice as long as the stories written by participants in either agent condition. However, neither difference was statistically significant. More specifically, using a double-sided t-test, we obtained $p = 0.062$ when comparing the baseline and story conditions and $p = 0.057$ when comparing the baseline and reader conditions.

Typing speed: The standard way to measure typing speed is as the average number of 'words' typed in a minute. Here, a 'word' refers to any 5 characters (including spaces), typed by the user.

We found that participants in the control condition, on average, typed approximately twice as fast as participants in either agent condition. Furthermore, using a double-sided t-test, we found that these differences were statistically significant. We obtained $p = 0.01$ and $p = 0.02$ when comparing the baseline with the story and reader conditions, respectively.

Number of hesitations: We defined a hesitation as any pause in typing that lasts 5 or more seconds. On average, participants in the reader condition were less hesitant than participants in the baseline condition, who were in turn less hesitant than the participants in the story condition. However, none of these differences were found to be statistically significant (all three p-values were greater than 0.31).

Average hesitation length: For each participant, the average length of their hesitations (as defined above), was measured in seconds. We found that, on average, participants in the baseline condition had hesitations that were approximately half the length of hesitations experienced by participants in the two agent conditions. Furthermore, using a double-sided t-test, these differences between the baseline and story conditions were found to be statistically significant ($p = 0.013$). However, when comparing the baseline and the reader conditions, we obtained $p = 0.051$, implying that the differences in these two conditions are not statistically significant.

Number of study pauses: Since we conducted our study asynchronously, our system had a 'pause' button to allow users to pause the study timer in case of an external distraction. We did not find any differences between the average number of times this feature was used by participants in the three conditions.

Table 5: Automated Measures of Writer's Block

Measure	Story	Baseline	Reader
final word count	Mean = 154.75 (SD = 63.05)	**Mean = 301.5** (SD = 104.5)	Mean = 150.75 (SD = 61.42)
typing speed (WPM)	Mean = 21.64 (SD = 5.29)	**Mean = 40.10** (SD = 8.35)	Mean = 23.42 (SD = 4.04)
num hesitat-ions	Mean = 12.50 (SD = 3.87)	Mean = 10.75 (SD = 4.11)	**Mean = 9.50** (SD = 3.70)
avg hesitat-ion len (sec)	Mean = 14.7 (SD = 3.09)	**Mean = 7.23** (SD = 0.68)	Mean = 12.66 (SD = 3.52)
num study pauses	Mean = 2.25 (SD = 2.63)	Mean = 2.25 (SD = 1.26)	Mean = 2.25 (SD = 2.06)

5.2.3 Qualitative Survey Responses.

To get a more nuanced understanding of common reasons for getting blocked in the writing process, we asked our participants to describe a time when they got stuck while writing their story. By coding the responses, we found that participants got stuck when they were: 1) lacking ideas, 2) trying to convey their ideas in words, 3) thinking about the time constraint, or 4) trying to relate their story to the provided topic.

Lack of ideas: 5 out of 12 of our participants implied that they got stuck during the writing task because they were out of ideas. The areas that these participants struggled with comprised thinking of a storyline and story events (3 participants), an ending (1 participant), or a setting (1 participant).

Conveying ideas in words: The next most commonly cited cause for getting stuck was an inability to put ideas into words (4 out of 12 participants). For example, one participant mentioned that they had a picture in mind but did not know to express it in a clear and interesting way.

Time-constraint anxiety: 2 participants mentioned that they got stuck after thinking about or being reminded of the study's time constraint.

Starting topic: 2 participants stated that they got stuck because they did not know how to relate their story to the starting topic provided in the study instructions.

Interestingly, out of the 5 participants who expressed difficulty in starting their story, all were from the two Prompter conditions. Since our system required participants to write at least 20 characters before getting help, this difference is likely either a sign that the assignment of participants to the different conditions was confounded, or that the 'help' button had an unintended effect on users' ability to start their story.

5.3 Usefulness of Different Prompts

Participants in the two agent conditions rated the usefulness of the prompts on a 7-point Likert scale (1: strongly disagree, 7: strongly agree). Figure 3 (Q14-Q20) shows the frequency distributions of responses compared across both conditions. As seen in Table 6, from the medians of the responses between the two conditions we observed that participants who received story-based prompts found the prompts more useful compared to participants who received reader-based prompts. However, using a two-sided Mann-Whitney U test to compare the results from each of the questions, we did not find that the differences between the two conditions were statistically significant (all six p-values were greater than 0.17). Nonetheless, it is promising that all participants who received story-based prompts at least somewhat agreed that the prompts helped improve their stories and were not distracting.

Table 6: Survey Results on Prompt Usefulness

Question	Story	Reader
Q14: helpful	**Med = 4.0** (IQR = 2.5)	Med = 3.0 (IQR = 2.5)
Q15: easy to incorporate	**Med = 4.5** (IQR = 2.5)	Med = 2.5 (IQR = 2.5)
Q16: relevant	**Med = 3.5** (IQR = 3.5)	Med = 2.0 (IQR = 1.5)
Q17: improved story	**Med = 5.5** (IQR = 1.0)	Med = 3.0 (IQR = 3.0)
Q18: better characters	**Med = 5.0** (IQR = 1.5)	Med = 3.5 (IQR = 2.0)
Q19: distracting	**Med = 2.5** (IQR = 1.0)	Med = 4.0 (IQR = 2.5)
Q20: not useful	**Med = 3.0** (IQR = 1.5)	Med = 4.0 (IQR = 2.5)

To compare the usefulness of the two prompt types, we also looked at log data of users' interactions with the system. Table 7 shows descriptive statistics comparing participant behaviour across the two agent conditions. We observed that, on average, Reader Prompter participants requested more prompts; however, they tended to be less satisfied with the prompts they received. This is evidenced by the fact that, on average, they left fewer stars on prompts as well as fewer and shorter notes when compared to the Story Prompter participants. However, running a double-sided t-test, we did not observed any statistically significant differences between the means of these four measures across the two conditions (all four p-values were greater than 0.29).

Table 7: Prompt Activity Per Participant

Measure	Story	Reader
number of prompt requests	Mean = 3.50 (SD = 1.29)	**Mean = 3.75** (SD = 3.30)
number of starred prompts	**Mean = 1.00** (SD = 1.15)	Mean = 0.25 (SD = 0.50)
number of prompts with notes	**Mean = 1.5** (SD = 1.29)	Mean = 0.5 (SD = 1.00)
number of words brainstormed	**Mean = 9.50** (SD = 8.43)	Mean = 3.25 (SD = 6.50)

The differences in these initial findings suggest that the wording of the prompts is likely an important design decision which should be investigated further.

5.3.1 Qualitative Data.
We asked the 8 participants who experienced Prompter to describe a time when a prompt was useful to them and why. 4 of these participants stated that the prompts helped them further develop their characters. For example, one participant mentioned that a prompt led them to add a second character and think about the characters' upbringings (*"I think it mentioned something about upbringing so I added a golden child and some least-favourite-child childhood trauma to the story"*). Another participant described how a prompt triggered a chain of ideas that helped make the main character more detailed (*"The ... prompt asking how the character's personality affected his decision led me to ... portray my character James as an introverted child. This led me to further ... ideas on how his changing to new schools every few months affected him and may have caused him to be introverted"*). Interestingly, 3 out of 4 of the participants who mentioned character development as a benefit of the prompts belonged to the story condition.

To get feedback on improving the prompts, we also asked participants to describe a time when a prompt did not help them. We observed four main themes: participants viewed prompts to be unhelpful because they were irrelevant, too late, did not help with the storyline, or did not help with the starting topic. 5 of 8 participants mentioned that some prompts were not helpful because they somehow did not fit their story. Examples include when a prompt referred to an earlier part of the text or a concept that was not the focus of the story. Furthermore, 2 participants mentioned that the prompts were too late because they were asking about something they had already thought about. Additionally, 2 of the participants did not find the prompts helpful since they were looking for help with their storyline. Similarly, 1 participant mentioned that they needed help coming up with a 'bizarre event,' which none of the prompts helped with.

5.4 Usability and User Feedback

5.4.1 Usability.
Participants rated the usability of the system on a 7-point Likert scale (1: strongly disagree, 7: strongly agree). Figure 3 (Q5, Q6, Q11-13) shows the frequency distributions of responses compared across the three conditions. To assess usability, we asked questions about the following constructs: ease of use, enjoyment, and intention to use in the future. As seen in Table 8 (Q11-13), participants in the reader condition tended to view the system more favourably compared to participants in the other two conditions. This is surprising since as discussed in section 5.3, participants in this condition found the prompts less useful than participants in the other agent condition. However, using a two-sided Mann-Whitney U test, none of these differences were found to be statistically significant. Surprisingly, participants in the baseline condition, compared to participants in the agent conditions, reported that they had more fun writing the story, and that they wanted to continue writing stories in the future (Table 8 - Q5,6). However, none of these differences were statistically significant. It is still promising that, at least in the story condition, participants' responses for all questions in this category were higher than the Likert-scale midpoint value of 4.

Table 8: Usability Survey Results

Question	Story	Baseline	Reader
Q11: easy to use system	Med = 6.0 (IQR = 0.0)	Med = 6.0 (IQR = 0.5)	**Med = 6.5** (IQR = 1.0)
Q12: enjoyable to use system	Med = 5.5 (IQR = 1.0)	Med = 5.5 (IQR = 1.0)	**Med = 6.5** (IQR = 1.5)
Q13: would use system in the future	Med = 5.0 (IQR = 0.5)	Med = 5.0 (IQR = 3.0)	**Med = 5.5** (IQR = 2.0)
Q5: fun while writing story	Med = 6.0 (IQR = 1.0)	**Med = 6.5** (IQR = 1.5)	Med = 5.0 (IQR = 2.0)
Q6: would keep writing stories in the future	Med = 4.5 (IQR = 1.0)	**Med = 5.5** (IQR = 2.0)	Med = 4.0 (IQR = 2.5)

5.4.2 Qualitative Data.
Analyzing the free-form text user feedback, we identified six common themes for areas of improvement: three applying to all conditions, and three applying only to the agent conditions. Regardless of condition, participants expressed their desire for more time and/or a better mechanisms for displaying how much time is left; seeing example stories before writing their own; and receiving non-story specific writing suggestions on their word use and grammar. Participants in the two agent conditions wanted to see prompts that were more relevant to the content of their stories; prompts that suggested ideas as well as questions; and clarification on the role of the notes section tied to the prompts. Specifically, in terms of

more relevant prompts, one user mentioned that they wanted the prompts to focus on what they had just written instead of directing their attention to earlier parts of the text. For ideas that could be displayed along with the prompts, suggestions included character name and personality generators.

6 DESIGN IMPLICATION

6.1 Discussion

Our findings from this initial pilot study have given us several insights that will be valuable for future work in improving Prompter. Three of our most surprising findings were that story-based prompts were preferred to reader-based prompts; the study length and UI timer caused anxiety during the writing process; and participants in the two agent conditions experienced more writer's block and wrote less than participants in the baseline condition.

6.1.1 Preference for Story-Based Prompts.

When comparing the median response of participants in the reader condition to that of participants in the story condition, we observed that the former group was more satisfied with the quality of their stories and also found the system to be more useful. The preference for Story Prompter was also reflected in our log data of users' interactions with the system. Although these results are not statistically significant, this repeating trend points to the importance of future research in determining how best to phrase prompts aimed at helping beginner story writers. We believe there are several potential explanations for this difference in preference among the two prompt types. Because of their wording, story-based questions were on average shorter than reader-based questions and more straightforward. These differences may have introduced unintentional confounding; nonetheless, this difference highlights the importance of short, to-the-point prompts in creative writing assistant tools.

6.1.2 Time-Caused Anxiety.

We observed that, regardless of the study condition, several participants mentioned that they felt rushed to finish their stories or that the time pressure of the writing task caused them anxiety. We chose to set the length of the writing task to be 10 minutes, since we believed longer sessions would be fatiguing for our participants; however, from this initial feedback, it is clear that future studies must provide adequate writing time to users. In addition, several participants pointed to our mechanism for displaying the remaining study time as a cause of anxiety. One participant mentioned that after seeing the '1 minute left' reminder, they felt blocked and stopped writing. Another participant mentioned that they did not see our reminders because they were focused on the writing task. This is an issue that we plan to address in future studies.

6.1.3 Writer's Block When Using Prompter.

The design of prompter was based on pedagogical research on methods for reducing the occurrence of writer's block. These methods suggested the importance of brainstorming as well as systematic question-asking. We were therefore surprised to find that participants who used Prompter reported experiencing more writer's block compared to those who did not. Even more surprising given our small sample size, we found a statistically significant reduction in the typing speed of Prompter users. There are several possible explanations for these findings. First, although we provided a notes area to encourage brainstorming, some participants were confused by this feature and found that it detracted from the writing exercise. Additionally, as previously mentioned, several participants felt rushed to finish their stories, even those in the baseline condition. This time pressure was more pronounced for participants who used Prompter, since they had more features to interact with. This reduction in the time available to write is a likely reason for the reduced word count and typing speed of Prompter users.

Of course, since it is difficult to control for every confounder when assigning participants to different conditions, the abovementioned findings may also be a result of unforeseen differences in demographics across the two groups. As a result, we plan on using a within-subjects experimental methodology in future studies, which will allow us to control for personal characteristics across the different conditions.

6.2 Future Work

6.2.1 The Role of Prompter in the Creative Writing Process.

In future work, we plan to explore the role of Prompter in the creative writing process. The literature on computational creativity points to the following components that must be taken into consideration when evaluating computational creativity tools: person, process, and product [21]. In the case of Prompter, we are particularly interested in its effects on the person writing the story as well as the process taken to write the story. More specifically, we wish to promote users' desires to pursue creative writing as a hobby and to scaffold their learning experience by giving them ideas when they are stuck. On the other hand, we are less interested in whether the final product is high-quality. These considerations will be important in helping us clarify our research questions. Furthermore, it is not clear at which stage of the writing process Prompter is more useful. In this study, we explored our tool in writing the first draft of a story, but we plan to further investigate the role of Prompter in the brainstorming and/or revision stages of writing.

6.2.2 Improvements to Prompter.

The feedback we received from users emphasized the importance of making Prompter more intelligent. As a result we plan on further investigating methods to improve the relevance and variety of prompts. For relevance, potential avenues for improvement may include only suggesting prompts for recently written text as well as replacing the Stanford CoreNLP library with state-of-the-art natural language classifiers such as BERT [11]. To increase the variety of prompts, we could introduce prompts about different story-elements such as the setting and plot, or use state-of-the-art text generators such as GPT-2 [33].

7 CONCLUSION

In this work, we designed and developed Prompter, an agent that asks students personalized questions based on their written stories. Through a survey of the literature on creative writing, we identified two broad categories of questions: story-based and reader-based. We conducted a pilot experiment with 12 participants, who received either no questions, story-based questions, or reader-based questions during their writing. Preliminary results show that Prompter's questions led participants to expand on their stories and that story-based questions were preferred to reader-based one. Future work in this area should explore the role of such question-asking agents in different stages of the writing process as well as technical improvements to make the agent more intelligent.

REFERENCES

[1] Mehdi Alaimi, Edith Law, Kevin Daniel Pantasdo, Pierre-Yves Oudeyer, and Hélène Sauzeon. 2020. Pedagogical agents for fostering question-asking skills in children. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[2] Michael Mose Biskjaer, Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, and Peter Dalsgaard. 2019. How Time Constraints in a Creativity Support Tool Affect the Creative Writing Experience. In *Proceedings of the 31st European Conference on Cognitive Ergonomics*. 100–107.

[3] Pietro Boscolo and Suzanne Hidi. 2007. The multiple meanings of motivation to write. *Writing and motivation* 19, 1 (2007).

[4] Deborah Brandt. 2009. *Literacy and learning: Reflections on writing, reading, and society*. John Wiley & Sons.

[5] Chris R Brewin, Tim Dalgleish, and Stephen Joseph. 1996. A dual representation theory of posttraumatic stress disorder. *Psychological review* 103, 4 (1996), 670.

[6] Alan Brown. 2007. Writing for Children. *The Handbook of Creative Writing* (2007), 162–168.

[7] Orson Scott Card. 2011. *Elements of Fiction Writing - Characters Viewpoint: Proven advice and timeless techniques for creating compelling characters by an a ward-winning author*. Writer's Digest Books.

[8] Sharon Lynn Chu, Francis Quek, and Kumar Sridharamurthy. 2014. Ready... action! a performative authoring system for children to create animated stories. In *Proceedings of the 11th Conference on Advances in Computer Entertainment Technology*. 1–4.

[9] 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.

[10] Paul Dawson. 2003. Towards a new poetics in creative writing pedagogy. *Text* 7, 1 (2003), 20.

[11] 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).

[12] Gerd Fischer, JS Gero, and ML Maher. 1993. Creativity enhancing design environments. *Modeling creativity and knowledge-based creative design* (1993), 235–258.

[13] Pauline Carolyne Fortes and Abdellatif Tchantchane. 2010. Dealing with large classes: A real challenge. *Procedia-Social and Behavioral Sciences* 8 (2010), 272–280.

[14] Richard P Gabriel, Jilin Chen, and Jeffrey Nichols. 2015. InkWell: A Creative Writer's Creative Assistant. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 93–102.

[15] Franca Garzotto. 2014. Interactive storytelling for children: a survey. *International Journal of Arts and Technology* 7, 1 (2014), 5–16.

[16] 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.

[17] Frederica Gonçalves, Ana Caraban, Evangelos Karapanos, and Pedro Campos. 2017. What shall i write next? Subliminal and supraliminal priming as triggers for creative writing. In *Proceedings of the European Conference on Cognitive Ergonomics 2017*. 77–84.

[18] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.

[19] Ernest Hemingway. 2012/1964. *A Moveable Feast*. Vintage Classics.

[20] Chieh-Yang Huang, Shih-Hong Huang, and Ting-Hao Kenneth Huang. 2020. Heteroglossia: In-situ story ideation with the crowd. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[21] Anna Jordanous. 2016. Four PPPPerspectives on computational creativity in theory and in practice. *Connection Science* 28, 2 (2016), 194–216.

[22] Alecia Marie Magnifico. 2010. Getting Others' Perspectives: A Case Study of Creative Writing Environments and Mentorship. (2010).

[23] Enrique Manjavacas, Folgert Karsdorp, Ben Burtenshaw, and Mike Kestemont. 2017. Synthetic literature: Writing science fiction in a co-creative process. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*. 29–37.

[24] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.

[25] Raymond A Mar. 2004. The neuropsychology of narrative: Story comprehension, story production and their interrelation. *Neuropsychologia* 42, 10 (2004), 1414–1434.

[26] EA Markham. 2007. Reading, Writing and Teaching the Short Story. *The Handbook of Creative Writing* (2007), 95–108.

[27] Peggy J Miller. 2014. Personal storytelling in everyday life: Social and cultural perspectives. *Knowledge and Memory: the Real Story: Advances in Social Cognition, Volume VIII* (2014), 177.

[28] Lawrence J Oliver. 1982. Helping students overcome writer's block. *Journal of Reading* 26, 2 (1982), 162–168.

[29] Jackie Peck. 1989. Using storytelling to promote language and literacy development. *The Reading Teacher* 43, 2 (1989), 138–141.

[30] Jonas Frich Pedersen, Lindsay Macdonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the landscape of creativity support tools in hci. In *The 2019 ACM CHI Conference on Human Factors in Computing Systems (CHI'19) ACM Conference on Human Factors in Computing Systems. Association for Computing Machinery (ACM)*.

[31] James W Pennebaker and Janel D Seagal. 1999. Forming a story: The health benefits of narrative. *Journal of clinical psychology* 55, 10 (1999), 1243–1254.

[32] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082* (2020).

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

[34] Melissa Roemmele and Andrew S Gordon. 2018. Automated assistance for creative writing with an rnn language model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. 1–2.

[35] Jane Rogers. 2007. Introduction to the Novel. *The Handbook of Creative Writing* (2007), 116–125.

[36] Mike Rose. 1980. Rigid rules, inflexible plans, and the stifling of language: A cognitivist analysis of writer's block. *College composition and communication* 31, 4 (1980), 389–401.

[37] George Saunders, Gogol Nikolai Vasilevich, Anton Pavlovich Chekhov, Anton Pavlovich Chekhov, Anton Pavlovich Chekhov, Ivan Sergeevich Turgenev, Leo Tolstoy, and Leo Tolstoy. 2021. *A swim in a pond in the rain: in which four Russians give a master class on writing, reading, and life*. Random House.

[38] Ben Shneiderman. 2009. Creativity support tools: A grand challenge for HCI researchers. In *Engineering the User Interface*. Springer, 1–9.

[39] Joshua M Smyth. 1998. Written emotional expression: effect sizes, outcome types, and moderating variables. *Journal of consulting and clinical psychology* 66, 1 (1998), 174.

[40] Masanori Sugimoto, Toshitaka Ito, Tuan Ngoc Nguyen, and Shigenori Inagaki. 2009. GENTORO: a system for supporting children's storytelling using handheld projectors and a robot. In *Proceedings of the 8th International Conference on Interaction Design and Children*. 214–217.

[41] Mary Swander, Anna Leahy, and Mary Cantrell. 2007. Theories of creativity and creative writing pedagogy. *The handbook of creative writing* (2007), 11–23.

[42] Cristina Sylla, Clara Coutinho, and Pedro Branco. 2014. A digital manipulative for embodied "stage-narrative" creation. *Entertainment Computing* 5, 4 (2014), 495–507.

[43] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* 17, 3 (2020), 261–272.

[44] Jan Vom Brocke, Wolfgang Maaß, Peter Buxmann, Alexander Maedche, Jan Marco Leimeister, and Günter Pecht. 2018. Future work and enterprise systems. *Business & Information Systems Engineering* 60, 4 (2018), 357–366.

[45] Lev Vygotsky. 1978. Interaction between learning and development. *Readings on the development of children* 23, 3 (1978), 34–41.

[46] Lev Semenovich Vygotsky. 2004. Imagination and creativity in childhood. *Journal of Russian & East European Psychology* 42, 1 (2004), 7–97.

[47] Thiemo Wambsganss, Sebastian Guggisberg, and Matthias Soellner. 2021. ArgueBot: A Conversational Agent for Adaptive Argumentation Feedback. (2021).

[48] Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. AL: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[49] Robert S Wyer Jr. 2014. *Knowledge and Memory: The Real Story: Advances in Social Cognition, Volume VIII*. Psychology Press.

[50] Niloofar Zarei, Sharon Lynn Chu, Francis Quek, Nanjie'Jimmy' Rao, and Sarah Anne Brown. 2020. Investigating the Effects of Self-Avatars and Story-Relevant Avatars on Children's Creative Storytelling. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.