Enhancing Peer Review with AI-Powered Suggestion Generation Assistance: Investigating the Design Dynamics

Seyed Parsa Neshaei
seyed.neshaei@epfl.ch
EPFL
Lausanne, Switzerland

Roman Rietsche
roman.rietsche@bfh.ch
Bern University of Applied Sciences
Bern, Switzerland

Xiaotian Su
xiaotian.su@epfl.ch
EPFL
Lausanne, Switzerland

Thiemo Wambsganss
thiemo.wambsganss@bfh.ch
Bern University of Applied Sciences
Bern, Switzerland

ABSTRACT

While writing peer reviews resembles an important task in science, education, and large organizations, providing fruitful suggestions to peers is not a straightforward task, as different user interaction designs of text suggestion interfaces can have diverse effects on user behaviors when writing the review text. Generative language models might be able to support humans in formulating reviews with textual suggestions. Previous systems use two designs for providing text suggestions, but do not empirically evaluate them: *inline* and *list of suggestions*. To investigate the effects of embedding NLP text generation models in the two designs, we collected user requirements to implement `Hamta` as an example of assistants providing reviewers with text suggestions. Our experiment on comparing the two designs on 31 participants indicates that people using the *inline* interface provided longer reviews on average, while participants using the *list of suggestions* experienced more ease of use in using our tool. The results shed light on important design findings for embedding text generation models in user-centered assistants.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; • **Human-centered computing → Natural language interfaces**.

KEYWORDS

intelligent writing assistants, text generation, peer reviews, natural language processing, machine learning, generative models

ACM Reference Format:
Seyed Parsa Neshaei, Roman Rietsche, Xiaotian Su, and Thiemo Wambsganss. 2024. Enhancing Peer Review with AI-Powered Suggestion Generation Assistance: Investigating the Design Dynamics. In *29th International Conference on Intelligent User Interfaces (IUI '24), March 18–21, 2024, Greenville, SC, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3640543.3645169

IUI '24, March 18–21, 2024, Greenville, SC, USA

ACM ISBN 979-8-4007-0508-3/24/03
https://doi.org/10.1145/3640543.3645169

1 INTRODUCTION

Peer review[1] is the process in which people evaluate the contents produced by peers in terms of positive and negative aspects, as well as discussing opportunities for improvement [53]. Individuals in various domains such as science, education, software development, and business make use of peer reviews to evaluate the quality of their work. For example, in the domain of science, providing peer reviews is considered an important part of the scientific publishing process [3]. Modern educational settings include peer review writing in their curriculum, as providing feedback is considered an important skill to possess as a current or future employee [75, 110]. Peer review is also used in the code-reviewing process of software development to detect bugs and maintain coding standards [50]. In addition, in large organizations, peer reviewing have been found to be a useful method to provide feedback to colleagues in agile settings [7, 11, 79, 84].

Despite that, producing reviews of high quality that are helpful and valuable to the original author is a challenge that reviewers frequently encounter [83]. Novice reviewers (e.g., students) may not be competent in providing a high-quality review [18]. For example, their reviews might lack elaborated suggestions for how to improve the original content [96]. As such, the feedback provided by peers may not be perceived as meaningful as intended by the review recipients [45, 53]. In addition to the challenges above, individuals often have trouble coming up with the right words to express what they intend, as with creating any other piece of writing [32, 36]. While resources are available to act as guides on how to write better texts, writers are not necessarily familiar with them and have to stop while writing to look up the resources and find suitable expressions to use, which leads to distraction and potential high cognitive load [38].

To address these challenges, researchers across domains [57, 83, 105] have explored the possible methods of supporting people in writing peer reviews easier with the use of technology. Liu et al. [65], for instance, has proposed developing a system that enables users to provide better reviews by creating and asking questions. Other researchers have used methods based on Natural Language Processing (NLP) and Machine Learning (ML) to provide review writing support and analyze written reviews, mostly based on text classification methods [13, 103, 104].

[1] Also called peer evaluation, peer feedback, or peer assessment

However, with the latest advances in ML and NLP, by making use of large-scale language text generation models such as GPT-2 [81] and GPT-3 [15], there is the potential to drastically improve the quality of writing support. Text classification methods can only be applied after the text is written to the end and given to the model as inputs [89], but on the other hand, text generation models overcome this limitation. Text completion interfaces[2] are used in various domains, such as text editors and search engines, to assist users in formulating a correct and relevant input [63]. They have the benefit of providing support to the users in the middle of the writing process. However, until today, there is a lack of research on how to best incorporate text completion interfaces in the context of peer review writing support to assist reviewers in providing feedback.

To address this research gap, we investigated the use of text generation models in providing writing assistance by analyzing the designs of text completion interfaces in previous works. Based on existing literature and user interviews, we derived potential designs and selected two of the widely used designs in previous literature and implemented systems: first, providing *inline suggestions* in the text area, and second, showing a *list of suggestions* next to the text area. The two selected design approaches from previous works have fundamentally different assumptions on how the suggestions might be used. While the former provides the possibility to directly accept the suggestions and alter them in the text area afterward, the latter mandates the user to either copy parts of the suggestions into the text editing field, or take ideas from them. To the best of our knowledge, there is no existing research comparing both approaches in helping people to write more helpful peer reviews. Therefore, we aimed to explore the effects of the two conventional designs of text completion-enabled writing assistants in the specific domain of computer-supported peer review writing as an example downstream task. To do so, we instantiated the two designs in Hamta, our newly designed and implemented writing assistant. By designing Hamta, we aimed to find the answer to our research question: "What are the effects of generative peer review writing support on users' reviews, as well as their perception and behavior with a text completion writing support tool?".

To be able to measure the effects of text completion designs on writing reviews, we used a dataset for training our German GPT-2 model to provide humans with intelligent suggestions in the domain of German business peer reviews. We utilized a corpus of 11,925 peer reviews written by students in a business course at a Western European university over five years, previously collected by Wambsganss et al. [107]. Based on the data, we fine-tuned a German GPT-2 model that generates individual texts and thus provides users with suggestions on what content (e.g., words, sentences, or expressions) to add to their peer review. To design the interface and user interaction flow of Hamta, we interviewed six users with experience in writing peer reviews, to find user experience design rationales. We ensured they have all used at least one text completion interface in the past. We implemented two versions of Hamta, each using one of the two text completion interface design methods.

To measure the two designs (*inline* and *list of suggestions*), we conducted a natural experiment with 31 participants. In our experiment, we asked the participants to write a peer review on a business model concept (for which we showed an introductory video in the beginning) to an imaginary peer. We randomly assigned 18 people to use the *inline* design and the others (13 participants) to use the *list of suggestions* design. Our results indicate that the participants who used the *inline* design entered a significantly higher amount of words in their review. On the other hand, participants using the *list of suggestions* design reported significantly higher perceived ease of use. The results shed light on how humans perceive the two text completion designs differently in the task of writing peer reviews. It is important to note that we did not compare the *quality* of reviews across groups in the current study, and rather only analyzed perception metrics as well as review length as a quantitative measure; we leave the rest for future work.

Our research provides three contributions to the peer reviews, writing assistants, and text completion design interfaces research streams. First, based on the empirical results from our experiment, we provide novel insights into the embedding of text generation models in peer review writing assistants, by comparing two conventional text completion designs. Second, we provide user requirements for a text completion system in the domain of writing peer reviews, collected from semi-structured interviews with target users. Third, we design Hamta, one of the first writing assistants for peer reviews using text generation models, based on the collected user requirements, and evaluate how the users perceive it in our experiment. We believe user-centric review writing assistants using text generation models and based on our findings in this research could offer legitimate solutions for assisting users in writing peer reviews more effectively and easily in professional environments.

2 RELATED WORK AND CONCEPTUAL BACKGROUND

We base our work on the literature about peer reviews, previous works on NLP and ML models which provide generative suggestions, the previous research on user-centered systems providing text completion and generative suggestions, and the literature on human-computer interaction (HCI) and user experience (UX) design ideas for writing support systems.

2.1 Peer Reviewing in Different Domains

According to Nicol [74], providing a peer review (peer feedback) is defined as evaluating the submission of other people by producing a written text. The peer review task is relevant in various domains, such as business, science, and education.

In business, the aspects in which peer reviewing is relevant are twofold. First, peer feedback is used to communicate the evaluation of how employees in a given workplace perform their duties so that they can improve upon their working style and improve their performance [35]. An example of this class of constructive business peer reviews is the 360-degree feedback process, in which the coworkers, subordinates, and managers of employees, in addition to the employees themselves, evaluate the performance of their colleagues by providing feedback [4]. Second, peer reviews are also

[2] Also called autocomplete interfaces

known to help evaluate the business model concepts of peers, which is useful in teaching business communication courses [82].

In science, the process of providing feedback is conducted by the reviewers of peer reviewed journals and conferences. In scientific peer feedback, the reviewers are expected to provide critical evaluations of the manuscripts, mention the positive and negative aspects of the submitted paper, and give to-the-point suggestions for improvement [70]. Not only do the original authors of the manuscript benefit from receiving a review, but providing feedback also has beneficial impacts on the reviewers (e.g., in terms of improvements in their writing skills) [68]. Thus, strengthening the peer review procedures can help improve the use of science in the academic world among scholars.

Moreover, the peer review process is also used in educational settings [73]. As an example, peer reviewing is used in the context of peer assessment, an arrangement in which students evaluate and grade their peers' work by considering the amount, level, or quality of the outcomes of the learning process followed by their peers [95]. This procedure can be carried out both in a centralized manner in which the instructors collect the peer assessments and set the final grades of the students, or in a decentralized manner in which the students themselves set the grades of their peers directly [5]. Additionally, peer reviewing is used to provide feedback to students in massive online open courses (MOOCs) [1]. Students are encouraged to provide feedback to their peers, which helps them to improve their writing skills [112].

We specifically take note of the role of peer reviewing in the cognitive models and processes of writing [9]. Writing is not considered only a way to translate ideas into written text, but also tailoring it to the needs of the readers (e.g., providing positive and negative aspects, as well as suggestions for improvements, in peer reviews) [40]. Writing typically involves a complicated interplay of various processes, putting significant demands on the humans' possibly limited capacity. Therefore, effective strategies are needed to handle the writing process well [40, 46]. Previous works have also discussed the comparison of writing strategies between novice and more expert writers [40, 47]. In the current work, we specifically focus on novice peer review writers.

2.2 Generative Suggestions with NLP and ML

Due to the rise in the computational power of the systems in use in the world, there has been a sharp increase in the various types of NLP and ML models, and lots of attempts to enable computers to generate pieces of text [51]. A popular approach in the past had been using recurrent neural networks (RNNs). Several other models have also emerged from the idea of RNNs [14, 20, 78, 116]. Another class of deep learning models focusing on generative tasks is the class of Generative Adversarial Networks (GANs) [25, 30, 44]. The principal difference between GANs and conventional models is that GANs use adversarial methods to train the model [44, 51]. Although GANs have been used extensively in the domain of generating images [29, 41, 52, 90], there has been relatively less use of them in the domain of text generation [51].

An important point in the history of NLP models is the introduction of Transformers [98], which outperformed RNNs in several benchmarks [97]. Transformers have been used extensively to provide solutions to various NLP tasks, such as machine translation [60], text classification [66], and question answering [86]. In addition, Transformers have been used for generating text [15, 80, 81]. An example of Transformer models used in text generation is the *Generative Pre-trained Transformer* (GPT) class of models [80], later evolved into GPT-2 [81] and GPT-3 [15]. Researchers can fine-tune GPT models to generate texts in any domain, given a dataset of texts in that domain. As a result, GPT and Transformer models can generate texts in various domains, such as medical texts [6], text summaries [58], stories [34], patent claims [62], codes [27, 91], and research papers [94]. However, studies of the effects of text generation models in the domain of *peer reviews* are rare and not widely studied.

2.3 User-Centered Systems with Generative Suggestions

One of the principles of HCI is known as *recognition over recall*, which indicates that people recognize things they see better than when they recall them from their memories [55, 56]. It naturally follows that text completion interfaces can turn a recall problem into a recognition problem, as the suggestions are provided to the users, removing the need to recall the words they want to type. As a result, text completion is extensively used in websites and application platforms, such as search engines [67].

Per the literature, text generation and completion systems can be classified into three main categories:

A) Word Completion: These kinds of systems provide suggestions for completing the current word or expression the user is typing at the moment [12]. Providing such suggestions can especially prevent some common mistakes, such as misspellings [69, 71].

B) Next Word Prediction: These systems will provide the users with a list of the words that they are most likely going to type after the current last word. This approach to text completion is especially useful for users who are typing on their smartphones, as the text completion suggestions can help them enter text faster despite relatively small keyboard buttons, by choosing the words from the suggestions instead of typing them [8, 31, 43, 48].

C) Multiple Words Prediction: These types of systems, possibly by utilizing NLP models such as GPT [28] or RNNs [72], can be used to complete the current sentence or paragraph the user is typing by suggesting one *or* multiple words at once. This approach has been used to help users who are writing relatively long bodies of text, such as emails [72, 92], essays [33, 87], and code [27, 91][3].

Also, the text completion schemes mentioned above can be incorporated into user-centric tools by one of the following two principal design ideas discussed in previous works and implemented in existing systems:

I) List of Suggestions: In this approach, the system provides a list of all possible suggestions to the user, mostly in the format of a drop-down list or a uniquely placed area on the page, and possibly ranked based on their likelihood of appearance in the text [108]. In the list, the users can find the word they want to insert in their textual input [55, 93, 115]. An example of this kind of

[3] In the current work, we considered peer reviews as long texts, so we selected these types of systems for the design of our peer review writing assistant.

approach is the text completion feature in most search engines [67] or integrated development environments (IDEs) for coding [24].

II) Inline Suggestions: In this design idea, the most probable expression, word, or rest of the sentence to follow the currently typed text is displayed after the cursor, and the users can choose if they want to accept the suggestion (e.g., by pressing the Tab button on the keyboard, or a GUI[4] element). The suggestion can also be placed directly into the text area as normal text, and the users can remove it in case they dislike it. While this approach is less commonly discussed in scientific literature, it is increasingly used in relatively new applications, such as GitHub Copilot [27], the email writing interface of Gmail [17], and Jenni AI [54].

As far as we are aware, there is a lack of research on comparing and contrasting the two mentioned methods of assisting people to write peer feedback. As a result, an objective of our study is to investigate the effects of the two standard designs discussed above and how people interpret them in the context of computer-supported review writing. We believe the interaction of humans with novel large-scale language models forms a new class of system design, especially in collaborative settings. However, due to the only recent rise of these models [15, 80, 81], the literature stream on the design of novel interfaces based on these text generation models has not formed consistently yet. We aim to contribute to the literature stream of using large language models for text generation in user-centric interfaces by instantiating and evaluating the conventional designs of text completion interfaces in writing peer reviews.

2.4 Writing Support Tools and Assistants

One of the first attempts at providing a preliminary writing support tool dates back to the 1920s with the "cut-up" method (called "découpé" in French) [16] including the process of cutting up and then later rearranging the pieces of text to create a new text. Due to the increase in the variety of NLP models which depend on the continuously increasing computational power of computers, computer-supported intelligent writing assistants have risen in popularity and are now used extensively to support users in writing texts [42]. Some examples of intelligent writing assistants include Grammarly[5], Hemingway[6], WordTune[7], and Ginger[8]. These tools can provide suggestions to improve the text, in addition to detecting grammatical spelling errors and mistakes.

Planning, *Translating*, and *Reviewing* are defined as the three principal writing processes [37], and NLP-powered writing assistants can support users in a variety of ways with regards to each of those. In *planning*, writing assistants propose certain expressions, sentences, or structures to the user as ideas. An example of them is BunCho, an assistant for creative story writing in Japanese, which uses GPT-2 [81] to provide suggestions [76]. As another example, Clark et al. [21] developed a system to help users with writing slogans. These kinds of assistants can make use of generative models to support users by providing example sentences. In *translating*, writing assistants can provide annotated examples and suggestions from others to help with putting ideas into written text, such as IntroAssist which includes a checklist of best practices, highlighted text functionality, and annotated examples to guide users in writing help requests [49]. Finally, in *reviewing*, writing support assistants guide users on how they can improve their initial content to reach a higher-quality final text. An example is such a system is AL, which analyzes the text the user provides to the system and identifies the level of argumentativeness and persuasiveness of the text while providing insights to the user to further improve the content [102]. Another example is the work of Weber et al. [109] which helps users with writing legal case solutions. More specifically, previous researchers have explored how writing assistants can support users in providing peer reviews, by providing suggestions for local or global revisions. As an example, the system introduced by Yang [113] presents users with example texts and revision suggestions on certain grammatical expressions and structures.

Overall, while previous studies have explored methods and systems to support users in the peer reviewing process by providing revision suggestions, the combination of text generation models with user-centric writing tools to provide assistance to users in their *peer reviewing process* is rare from an HCI perspective. Moreover, there is a lack of studies comparing the usability of the various potential text completion user interfaces when embedded into a user-centric assistant for peer review writing support. We address this research gap by implementing a novel peer review writing assistant built around a GPT-2 text generation model and evaluating the various possible designs of such a system with a user study.

3 DESIGN OF GENERATIVE WRITING SUPPORT SYSTEM FOR PEER REVIEWS

To explore the effects of the design of user-centric writing assistants using text generation models and to compare the design methods we collected from previous works (see Section 2.3), we designed our novel generative writing tool[9]. We designed Hamta based on the design methods for text completion interfaces and writing assistants extracted from the suggestions proposed by the participants in our interviews. Designing our own tool enabled us to have a base text completion system in which we could conduct our experiment on using text generation models in peer review writing.

An overview of the architecture of our system can be seen in Figure 1. The users start writing their peer reviews. After typing in at least 25 words and waiting for a short time, the currently-typed review text is sent to the server and a GPT-2 model suggests three possible continuations based on the current text. The texts are returned to the front-end and displayed to the users as either *inline* suggestions or in a *list*, the two main design ideas for text completion interfaces which we extracted from previous works (see Section 2.3). In this section, we describe the procedure we followed to design Hamta, how the design rationales were instantiated in our tool for the users to interact with, and how we implemented the back-end of our system.

[4] Graphical User Interface

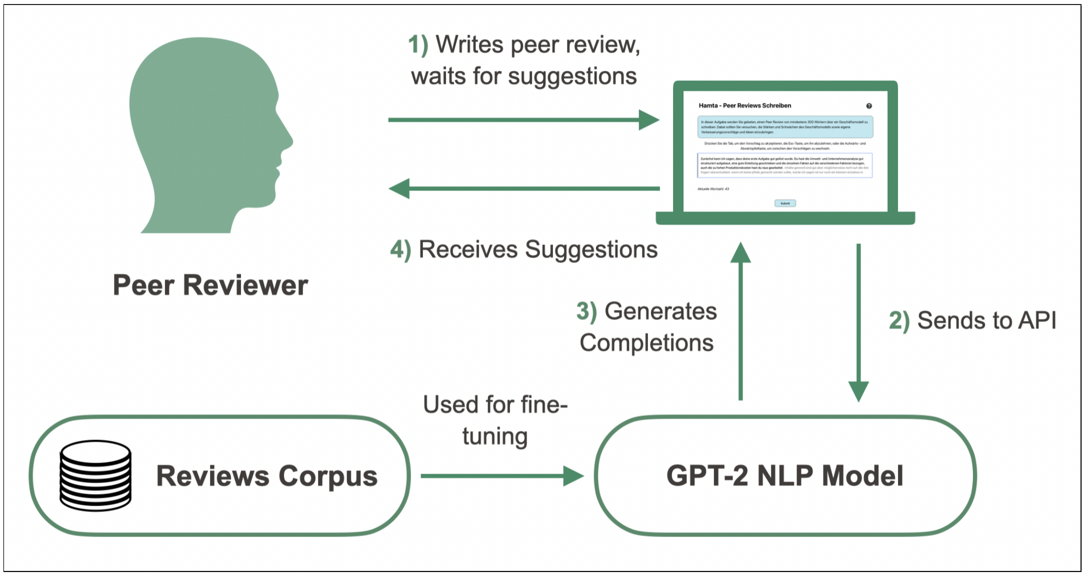
[5] https://grammarly.com

[6] https://hemingwayapp.com

[7] https://wordtune.com

[8] https://gingersoftware.com

[9] We translated the interface from German to English for demonstration in the figures of this paper.



Hamta - Peer Reviews Schreiben

In dieser Aufgabe werden Sie gebeten, einen Peer Review von mindestens 300 Wörtern über ein Geschäftsmodell zu schreiben. Dabei sollten Sie versuchen, die Stärken und Schwächen des Geschäftsmodells sowie eigene Verbesserungsvorschläge und Ideen einzubringen.

Drücken Sie die Tab, um den Vorschlag zu akzeptieren, die Esc-Taste, um ihn abzulehnen, oder die Aufwärts- und Abwärtspfeiltaste, um zwischen den Vorschlägen zu wechseln.

Zunächst kann ich sagen, dass deine erste Aufgabe gut gelöst wurde. Du hast die Umwelt- und Unternehmensanalyse gut strukturiert aufgebaut, eine gute Einleitung geschrieben und die einzelnen Fakten auf die verschiedenen Faktoren bezogen, auch die zu hohen Produktionskosten hast du raus gearbeitet . inhalte generell sind gut aber möglicherweise nicht auf alle drei fragen releventuellant. wenn ich keine pfeile gemacht werden sollte, würde ich sagen ist nur noch ein kleinem einzelnes in

Aktuelle Wortzahl: 43

Figure 1: An overview of the architecture of `Hamta`: The tool shows possible continuations of the current text, which the users can either use directly or take ideas from.

3.1 User Requirements

As also discussed in Section 1, a goal of our study is to help novice users write peer reviews, so we performed semi-structured interviews with six users as our target group. In this vein, we followed the suggestion by Cooper et al. [23] to engage with individuals referred to as "potential users," those who are not presently using the product but are considered promising candidates for future use. We specifically did not interview senior researchers or users with extensive experience in writing peer reviews, to see what a beginner, which matches our target group, would require from such a system. However, we still made sure that our interviewees had experience reviewing at least once and had used at least one text completion interface in a writing assistant in the past, to ensure they were not clueless about the interview topic. Among the interviewees, three were female and three were male.

Each interview lasted around 20 minutes. We asked the following topics in our interviews: *past experience with the peer review process*, *past experience with text completion interfaces*, *how they like to be supported in writing peer reviews using generative tools*, and *how and when they like the suggestions to appear*. Then, we showed an initial sketch of the two designs of our tool to the interviewees and asked them to comment on the design, along with providing suggestions on what they think is missing in the design or needs to be improved or changed. We showed the *inline* design to the first three interviewees and the *list of suggestions* design to the other three participants[10]. The interviewer took notes of all comments provided by the participants in the interview process.

To analyze the data from the semi-structured interviews, we followed the approach of thematic analysis [22]. Two researchers carefully annotated and classified the comments in the notes by the interviewer, and generated initial themes from the responses. They resolved any conflicts in the annotation process as they happened at the same time; thus, no inter-rater agreement was calculated.

From the responses of the interviewees, we found that they did not want to be forced to accept the suggestions, as based on their previous experience, the suggestions could be incorrect or misleading at times. They also preferred to see multiple suggestions as opposed to only one, as a single suggestion does not necessarily contain all the possible ideas on how to continue the text. We also asked for the number of suggestions they prefer to see. Four participants said exactly three, one said exactly two, and one said "two or three". Thus, we assume three suggestions are acceptable among the participants overall. They also mentioned the importance of returning to a part in the middle of the text and writing in the middle, and they said the suggestions should be relevant to the part of the text they are writing into. Regarding *when* they wanted to see the suggestion, they said they would like to receive suggestions after typing each word[11] and after a certain amount of time[12]. The participants also mentioned the wish to see how many words they have typed to get a sense of how much they have addressed the task of peer review they are undertaking.

Among the participants to whom we presented the sketch of the *inline* design of providing text suggestions, they mentioned the importance of the contrast in color between the text entered by the user and the suggestions added by the system, which was already incorporated in the initial sketch. In addition, they preferred using keyboard shortcuts as opposed to user interface buttons to accept,

[10] We specifically did not ask questions about *review quality* and related requirements, as we leave measuring review quality, rather than quantity and perception measures, for future work.

[11] As such, we assume the suggestions would appear after pressing the space key, which indicates the end of a word and/or sentence.

[12] Three of them specially mentioned "less than 10 seconds" for the preferred time of the timer which would provide the suggestions after it is fired.

reject, or move between the suggestions, as they did not want to be interrupted while typing by moving their hands from the keyboard to the mouse or trackpad. Among those who saw the sketch of the *list of suggestions* design, they preferred to see the suggestions in a fixed area on the right instead of a dropdown. They believed the dropdown design in such cases covered major parts of the text area, as the suggested texts in the domain of peer reviews consist of multiple words. They also mentioned the need for the possibility to copy parts of the displayed suggestions and paste them into the text area.

Finally, topic-wise clusters including the user stories were refined and finalized by the researchers, and thus we finally obtained nine clusters, indicating the user requirements below:

(U_1) The tool should not automatically insert the suggestions into the written content[13], but display them in a contrasting color (*inline* version) or in a separate section (*list of suggestions* version).

(U_2) Users should be presented with up to three suggestions, based on the review text written from the beginning up to the cursor's current position in the text area.

(U_3) Suggestions should only be generated once the user completes a word and/or a sentence (i.e., after pressing the Space bar on the keyboard).

(U_4) The tool should wait for a short period (ideally no longer than 10 seconds) before displaying text suggestions[14].

(U_5) The system should display the number of words entered by the user, updated in real-time.

(U_6) (Only applicable for the *inline* design) The suggestions should be displayed in a lighter color[15] than the original written review text.

(U_7) (Only applicable for the *inline* design) The system should include shortcuts[16] to accept, reject, or move through the suggestions[17].

(U_8) (Only applicable for the *list of suggestions* design) The suggestions should be presented in a list on the right side instead of being displayed in a dropdown to allow for a bigger space to see the suggestions in[18].

(U_9) (Only applicable for the *list of suggestions* design) The system should enable the functionality of copying parts of the presented suggestions and pasting them in the main text area.

3.2 User Interface of Hamta

Figure 2 shows the user interface consisting of six principal design functionalities (F_1 to F_6): the *inline suggestion* design (A) on the top, and the *list of suggestions* design (B) on the bottom[19].

Both interfaces include a text area (F_1) that users can use to type their reviews in. The text areas expand in height by entering more text or going to the next line by pressing Enter. The text area includes all standard functionalities of a conventional web text area, including copying and pasting (related to U_9). Users can also see the number of words they have written (F_2, related to U_5). After typing at least 25 words in their review, the system will become ready to provide suggestions. After typing each word (related to U_3), the text up to the current position of the cursor is sent to the back-end and then a GPT-2 model starts generating three suggestions (related to U_2). The system waits for at least 5 seconds and less than 10 seconds before showing the suggestions to the users (related to U_4). In the period the system is waiting before showing the suggestions, the text "Loading suggestions" will appear in the user interface (above the text area in the *inline* design, and above the *list of suggestions* in the other design).

After the suggestions are returned to the front-end, they are presented differently in the two designs. Neither of the designs inserts the suggestions as normal text automatically (related to U_1). In the *inline suggestions* design, the first of the three suggestions is appended in a lighter color after the current position of the cursor (F_3, related to U_6). As also evident by the prompt shown at this stage above the text area, users can press *Tab* to accept the current suggestion, press *Up* and *Down* arrow keys to move between the suggestions, or press *Esc* or any other alphanumerical key to reject the suggestion (related to U_7). In the *list of suggestions* design, the suggestions are presented in the box on the right side of the system (F_4) without blocking the effective text area (related to U_8). The box on the right side allows the suggestions to be copied with the standard system context menu or shortcuts so that they can be pasted into the main text area (related to U_9). In both designs, users can click the Submit button (F_5) to have their review saved in the back-end[20]. The users may also benefit from an always-available help button, which offers guidance if they get confused or become unsure how to progress (F_6, also backed by the literature [114]).

We implemented Hamta based on these functionalities and their link to the user requirements, as well as considerations we took from the literature (see Section 2.3). We built our tool as a React responsive web application, which does not mandate specific hardware requirements and can be used on a diverse range of devices (except smartphones). We aimed to provide an intuitive and simple-to-use interaction flow to the users so that they can start using the system to augment their writing skills without needing to see long guides or tutorials first.

[13] This ensures users are not mandated or pushed to accept the suggestions, which was mentioned in the user interviews.

[14] In our system, we assumed a fixed time of 5 seconds, in addition to the inference process which mostly takes less than 3 seconds.

[15] We chose a variation of the grey color in our system.

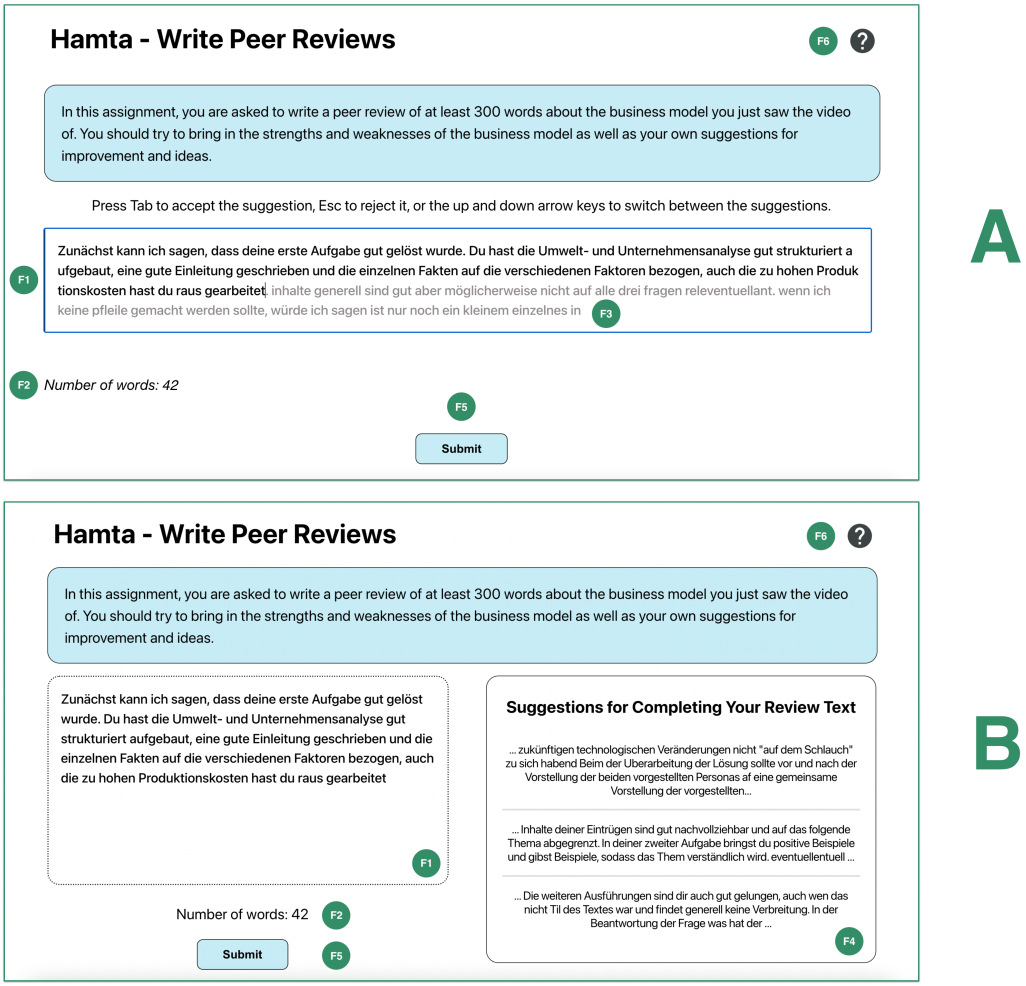
[16] We chose the *Tab* key to accept the suggestion, the *Esc* key or any alphanumerical key to reject it, and *Up* and *Down* arrow keys to move through the suggestions.

[17] We initially guessed that this would lead to writing a higher number of words when using the *inline* design, as accepting the whole suggestion at once would be easier for users. This speculation was also later confirmed in our experiment.

[18] We initially guessed that this would help with a higher perceived ease of use when using the *list of suggestions* design, also later confirmed in our experiment.

[19] While we extended and complemented the preliminary recent work of Su et al. [88] in our work, we A) launched a new round of user interviews from the ground up to inform the design of our tool more rigorously by considering two different designs, B) implemented our collected user requirements across two different design versions of Hamta, and C) focused the evaluation of our system, by comparing the two different design modalities of Hamta, on target users from Prolific rather than students in higher education, and thus, did not limit our system to only specific educational scenarios.

[20] This was especially needed for our experiment, as we aimed to collect the reviews of the users to conduct quantitative studies on them.



Hamta - Write Peer Reviews

In this assignment, you are asked to write a peer review of at least 300 words about the business model you just saw the video of. You should try to bring in the strengths and weaknesses of the business model as well as your own suggestions for improvement and ideas.

Press Tab to accept the suggestion, Esc to reject it, or the up and down arrow keys to switch between the suggestions.

Zunächst kann ich sagen, dass deine erste Aufgabe gut gelöst wurde. Du hast die Umwelt- und Unternehmensanalyse gut strukturiert a
ufgebaut, eine gute Einleitung geschrieben und die einzelnen Fakten auf die verschiedenen Faktoren bezogen, auch die zu hohen Produk
tionskosten hast du raus gearbeitet. inhalte generell sind gut aber möglicherweise nicht auf alle drei fragen releventuellant. wenn ich keine pfleile gemacht werden sollte, würde ich sagen ist nur noch ein kleinem einzelnes in

Number of words: 42

Hamta - Write Peer Reviews

In this assignment, you are asked to write a peer review of at least 300 words about the business model you just saw the video of. You should try to bring in the strengths and weaknesses of the business model as well as your own suggestions for improvement and ideas.

Zunächst kann ich sagen, dass deine erste Aufgabe gut gelöst wurde. Du hast die Umwelt- und Unternehmensanalyse gut strukturiert aufgebaut, eine gute Einleitung geschrieben und die einzelnen Fakten auf die verschiedenen Faktoren bezogen, auch die zu hohen Produktionskosten hast du raus gearbeitet

Number of words: 42

Suggestions for Completing Your Review Text

... zukünftigen technologischen Veränderungen nicht "auf dem Schlauch" zu sich habend Beim der Uberarbeitung der Lösung sollte vor und nach der Vorstellung der beiden vorgestellten Personas af eine gemeinsame Vorstellung der vorgestellten...

... Inhalte deiner Einträgen sind gut nachvollziehbar und auf das folgende Thema abgegrenzt. In deiner zweiter Aufgabe bringst du positive Beispiele und gibst Beispiele, sodass das Them verständlich wird. eventuellentuell ...

... Die weiteren Ausführungen sind dir auch gut gelungen, auch wen das nicht Til des Textes war und findet generell keine Verbreitung. In der Beantwortung der Frage was hat der ...

Figure 2: Screenshots of the two interfaces for Hamta, our generative writing assistant developed to explore the effects of using text generation models to support users in writing peer reviews. Users write feedback on a business project of a peer and receive possible words, sentences, or expressions to continue their text with, either inline (A) or in a list on the right side (B).

3.3 Back-end Algorithm of Hamta

We developed the back-end of our tool using Flask[21] in Python. This back-end is responsible for A) collecting the review texts the users have submitted, B) predicting multiple tokens as suggestions to continue the text using GPT-2, and C) saving the intermediate results in log files for further inspection.

3.3.1 Corpus. To fine-tune the German GPT-2 text generation model, we aimed to find a corpus providing enough text bodies so that our model can be trained on with acceptable results [39, 61]. We planned to choose a corpus from the previous works with the following attributes: A) the texts in the corpus are previously written "peer review"s rather than general-domain texts, to make the training context as close as possible to the inference context, and B) the number of sentences in the corpus is high enough, as text generation models naturally need a lot of data in their training process to perform well in practice. We finally selected a corpus of 11,925 peer reviews, all written by university students in a business department course at a Western European university [107]. The data was gathered collectively by researchers over five years, and has been previously used for investigating biases in educational writings [106]. This corpus enabled us to fine-tune our GPT-2 model to generate texts in the domain of business peer reviews, in line with what we aimed Hamta should be skillful at.

3.3.2 Training the model. Before using the dataset as it is to fine-tune our GPT-2 model, we initiated a data-cleaning phase to provide only clean data as the training input to our model. For example, specific information (e.g., URLs, specific names of business entities, and the prompt question asking the students to write reviews based upon, which some of them had copied to their reviews) were removed from the texts.

[21] https://flask.palletsprojects.com

It is important to note that the present version of our tool can only offer text generation assistance for reviews written in German because the dataset we used only contains German texts [107]. Hamta will be able to offer writing support in more languages, though, if similar datasets are added in additional languages.

4 EXPERIMENTAL EVALUATION

In this section, we will investigate the peer reviews and how users experienced the generative review writing support system we designed based on our collected user requirements, Hamta, to answer our research question and compare the two text completion interfaces. We conducted our field experiment over Prolific[22], a crowdsourcing platform for experiments. We selected Prolific because previous studies on behavioral research platforms found that Prolific had the highest response quality and sample variety [77], crucial criteria for evaluating crowdsourcing platforms [19, 85, 111].

We aimed to conduct the experiment among individuals who can resemble the end users of our study the most. As the target group for our system includes literate users who write projects or business feedback in professional environments and workplaces, we set the eligibility criteria of our study by only conducting the experiment on people who completed at least a high-school diploma and were previously or are currently employed, by setting the relevant filters in the Prolific panel when defining the study online. This ensured that our experiment would support claims on the helpfulness of Hamta for our desired target user group. Moreover, as the current version of Hamta includes a GPT-2 model which is only fine-tuned on German peer reviews [107], it could only generate suggestions in German. As a result, we only performed the experiment on participants who were fluent in German. The filters set in Prolific ensured that the participants in our experiment were viable "peer"s in our peer review process. Out of the 31 participants in our experiments, 14 were female and 17 were male (mean age = 35.0, SD = 14.9).

In our experiment, we asked the participants to start writing a review of the business pitch of an imaginary peer. To draw causal inferences, we randomly assigned participants to use one of the two designs introduced in Section 2.3: *inline suggestions* and *list of suggestions*[23]. From the 31 participants in total, 18 participants used the *inline suggestions* design (group A), while the rest (13 participants) used the *list of suggestions* design (group B). In group A (*inline suggestions* design), 8 participants were female and 10 were male (mean age = 38.2, SD = 17.6). In group B (*list of suggestions* design), 6 participants were female and 7 were male (mean age = 30.6, SD = 8.8). We granted each participant around 4.9 US dollars for participating in our experiment and filling out our questionnaire for around 30 minutes.

4.1 Experiment Design

As also discussed in Section 1, the main research question (**RQ**) we aim to provide an answer to is "What are the effects of generative peer review writing support on users' reviews, as well as their perception and behavior with a text completion writing support tool?". We designed our experiment with regard to our RQ. Our study contained a pre-survey, a main task (writing a peer review using one of the two designs), and a post-survey. Apart from the two designs (*inline suggestions* for group A and *list of suggestions* for group B), the other elements of the survey, including all the questions, were exactly the same among all the participants.

4.1.1 Pre-survey. As we assigned the users randomly to the two groups in our experiment, we designed our pre-survey to make sure participants were distributed randomly and there were no significant differences in technology usage and feedback seeking among them. We used two constructs in the pre-survey: A) *information technology usage model* [2], and B) *feedback seeking* behavior of participants [10]. We chose construct A to assess users' attitudes and behaviors concerning new technology, because we deemed it crucial to evaluate how prepared they are to adopt innovative platforms and tools, like Hamta. The items in construct A included *"I like to experiment with new information technologies," "If I heard about a new information technology, I would look for ways to experiment with it," "In general, I am hesitant to try out new information technologies,"* and *"Among my peers, I am usually the first to try out new information technologies"*. The second construct was selected to measure the extent to which participants sought feedback and to examine participants' attitudes regarding receiving suggestions, which we deemed crucial when assessing the impact of our writing assistant. The items in construct B included *"It is important for me to receive feedback on my performance"* and *"I find feedback on my performance useful"*. We measured both constructs with a 7-point Likert scale [64], anchored at 1 for complete disagreement, 7 for complete agreement, and 4 for being neutral. The two pre-survey constructs we chose have been employed in prior studies focusing on user-centric systems [101, 102] to check for valid randomization.

4.1.2 Writing Peer Review Task. After the pre-survey, we showed to the participants a three-minute video on a business model pitch, which they had to provide a review for. The video introduced the business model of a new application idea for scheduling trips with friends, as well as finding resorts to visit. The video also described the business model of the app and provided details on how the developers of the platform would charge fees from the users to keep the app running. Then, we gave the link to the respective design version of our tool to the participants and asked them to evaluate the business pitch in the video by writing a peer review. Participants were free to use suggestions as much as they wanted. We asked the participants to press the Submit button in the user interface at the end to save the review text. Specifically, the number of words participants submitted in their reviews was also calculated and stored in our back-end for further investigation. It is important to note that we did *not* consider the *word count* as an indicator of the review quality, and only reported it as a quantitative measure; we leave investigating the review *quality* measures separately for future work.

4.1.3 Post-survey. After the participants finished writing their review and submitted it, they entered the post-survey, which intended to ask the participants to reflect on their reviewing process and interaction with our tool, Hamta. The post-survey started with an "attention check" question, in which we asked the participants to

[22] https://www.prolific.co

[23] We did not also conduct a within-subject comparison and only conducted a between-subject evaluation, i.e., no participant tried both of the systems, due to budget availability for our research. We leave a within-subject comparison to future work.

select the word we asked them to remember in the video in the pre-survey. The results and responses from the participants who failed the attention check were excluded from the reported results in this paper. We selected the constructs to measure inspired by what has been used and validated in the literature [99, 100], as well as prior works on the evaluation of user-centric systems [101, 102]. We have included a list of the used constructs, as well as example questions asked in each of them in our post-survey, in Table 2 in the Appendix. It is important to note that as we frame our work around the concept of user-centric systems and writing assistants in general, we chose constructs used in HCI research for evaluating systems rather than utilizing validated questionnaires for review writing in particular. We leave investigating a more diverse set of questions in our pre- and post-survey question sets for future work.

5 RESULTS

We first calculated the average of the responses to the questions in the pre-survey constructs for each of the participants to be sure our randomization was valid. We then compared the responses among the participants in the two groups using the Wilcoxon rank-sum test. The Wilcoxon tests led to $p = .374, d = 0.086$ for the *information technology usage model* [2] and $p = .359, d = 0.236$ for the *feedback seeking* [10] construct[24]. Consequently, we found no difference between the two groups regarding the pre-survey questions. Thus, we consider the comparison in our study and the way we separated the participants into the two groups as valid.

5.1 Impact of Text Generation Designs on Users Perception

We performed a group comparison on the construct means using the Wilcoxon rank-sum test for our two designs (discussed in Section 2.3): A) *inline suggestions* and B) *list of suggestions*.

- **Technology Acceptance:** We obtained (M = 4.1, SD = 1.7) for group A, and (M = 4.2, SD = 1.6) for group B ($p = .476, d = 0.230$), so while the mean of group B is slightly more than that of group A, there is no significant difference between the two groups in terms of our statistical test. Both means are more than the baseline of 4.0.
- **Perceived Usefulness:** We obtained (M = 4.0, SD = 1.6) for group A, and (M = 4.7, SD = 1.7) for group B ($p = .107, d = 0.676$), so while the mean of group B is more than that of group A, there is no significant difference between the two groups in terms of our statistical test. Both means are more than the baseline of 4.0.
- **Perceived Review Quality:** We obtained (M = 4.6, SD = 1.3) for group A, and (M = 4.8, SD = 1.3) for group B ($p = .352, d = 0.319$), so while the mean of group B is slightly more than that of group A, there is no significant difference between the two groups in terms of our statistical test. Both means are more than the baseline of 4.00.
- **Perceived Ease of Use:** We obtained (M = 4.6, SD = 1.4) for group A, and (M = 5.4, SD = 1.6) for group B ($p = .044 < .05, d = 0.508$), so the difference between the two groups in this construct is **significant**. This shows the participants in group B (using the *list of suggestions* design) had a more comfortable interaction experience when using the system. Both means are more than the baseline of 4.00.
- **Perceived Improvement in Writing Reviews:** We obtained (M = 4.0, SD = 1.6) for group A, and (M = 4.5, SD = 1.7) for group B ($p = .218, d = 0.540$), so while the mean of group B is more than that of group A, there is no significant difference between the two groups in terms of our statistical test. While the mean of group B is more than the baseline of 4.00, this is not the case for group A.
- **Perceived Improvement in the Reviews in the Long Run:** We obtained (M = 4.2, SD = 1.8) for group A, and (M = 4.6, SD = 1.7) for group B ($p = .315, d = 0.468$), so while the mean of group B is slightly more than that of group A, there is no significant difference between the two groups in terms of our statistical test. Both means are more than the baseline of 4.00.
- **Correctness of the Suggestions:** We obtained (M = 4.2, SD = 1.8) for group A, and (M = 3.9, SD = 1.7) for group B ($p = .254, d = 0.016$), so while the mean of group A is slightly more than that of group B, there is no significant difference between the two groups in terms of our statistical test. While the mean of group A is more than the baseline of 4.00, this is not the case for group B.
- **Excitement After Interaction:** We obtained (M = 4.3, SD = 1.6) for group A, and (M = 4.3, SD = 1.8) for group B ($p = .429, d = 0.182$), so while the mean of group B is slightly more than that of group A, there is no significant difference between the two groups in terms of our statistical test. Both means are more than the baseline of 4.00.

A summary of conducting the Wilcoxon tests is included in Table 1. Also, the Cronbach's alpha [26] for all the eight constructs was more than or equal to 0.9 when rounded to one decimal place, indicating a high reliability of our questionnaire.

5.2 Quantitative Impact of Text Generation Designs on the Reviews

We also measured the number of words the users had written in the reviews among the two groups as a metric indicating the quantity of the text and compared them to find their difference statistically by conducting the Wilcoxon tests. We obtained (M = 318.1, SD = 76.5) words in group A (using the *inline suggestions* design) and (M = 261.8, SD = 103.0) words in group B (using the *list of suggestions* design), with $p = .029 < .05, d = 0.744$, so the difference between the two groups in this construct is statistically **significant**. This shows the participants in group A (using the *inline suggestions* design) entered a considerably higher number of words (including the accepted suggestions, as well as the words typed by the users themselves) in their reviews[25].

Again, we included an overview of conducting the Wilcoxon tests in Table 1.

[24] d = Cohen's d

[25] It should be noted that we did not consider this quantitative measure as an indication of review *quality*, but rather leave exploring quality of the outcome reviews for future researchers.

Table 1: Mean, standard deviation, and the one-sided p-value of conducting *Wilcoxon test*s on various constructs of metrics on the results of the survey and the reviews among the two groups: A with *inline suggestions* design, and B with *list of suggestions* design. All constructs (except *number of words*) are measured with the Likert scale (1: low, 7: high). $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

Metric	Mean A	Mean B	SD A	SD B	p-value
Information Technology Usage Model [2]	5.4	5.4	0.8	1.2	.374
Feedback Seeking [10]	5.3	5.4	0.7	0.8	.359
Technology Acceptance	4.1	4.2	1.7	1.6	.476
Perceived Usefulness	4.0	4.7	1.6	1.7	.107
Perceived Review Quality	4.6	4.8	1.3	1.3	.352
Perceived Ease of Use	4.6	5.4	1.4	1.6	**.044***
Perceived Improvement in Writing Reviews	4.0	4.5	1.6	1.7	.218
Perceived Improvement in the Reviews in the Long Run	4.2	4.6	1.8	1.7	.315
Correctness of the Suggestions	4.2	3.9	1.8	1.7	.254
Excitement After Interaction	4.3	4.3	1.6	1.8	.429
Number of Words in the Reviews	318.1	261.8	76.5	103.0	**.029***

6 DISCUSSION

In this research, we explored the impact of the possible designs of professional peer review writing support using text generation models by answering our research question (as discussed in Section 4.1). To do so, we first investigated the past works on user-centric text completion interface designs and extracted two main design interfaces from previously implemented systems, which were not previously evaluated in this context. Then, to be able to evaluate the two designs and investigate their effects on peer reviews, we embedded them in our user-centric tool, *Hamta*. To build our tool, we used a previously collected dataset [107] to fine-tune our GPT-2 model. To achieve the best results in terms of usability and ease of use metrics, we collected user experience design rationales from our interviews with target users to embed in the user interface of our tool.

The empirical results of our group comparison indicate that both designs (*inline* and *list of suggestions*) make people engaged when they interact with them. Moreover, both designs are considered to be useful and accepted among the participants. Additionally, limiting the scope of our experiment to participants resembling our target users, as well as conducting user interviews with members of our target group, ensured that `Hamta` could meet the needs of the potential target users. Consequently, the findings from our user study offer highly encouraging outcomes for using writing assistants designed based on our collected user requirements in a variety of professional review writing situations.

When comparing the results of the *inline suggestions* group and the *list of suggestions* group, we found a significant difference in *perceived ease of use* and *number of words*. The results indicate users were more comfortable working with the *list of suggestions* design. This confirmed our original speculation that in the *list of suggestions* design, they were able to see multiple suggestions at once in a designated area on the right side (as opposed to having to move between them using the arrow keys in the *inline suggestions* design), and thus they were able to choose among them quicker and with more comfort. We also observe a greater mean of the *number of words* in the *inline suggestions* design. This also confirmed our initial guess that in the *inline suggestions* design, accepting the whole suggestion text is very easy (with a single press of the Tab key on the keyboard). On the other hand, in the *list of suggestions* design, users may first see and compare the suggestions, and then copy or use the ideas in the texts instead of accepting the whole text all at once. Thus, we assumed accepting the whole suggestions all at once leads to writing more words in the same period for the review writing task, attributing to this significant difference in the results of our experiment. To summarize, we can conclude while the users are more comfortable working with the *list of suggestions* design, the *inline suggestions* design helps them write more words at the same time.

6.1 Theoretical and Practical Contributions

Previous studies looked into the systems and their effects on helping users write peer reviews more efficiently [83, 103, 104, 107, 113]. These tools provide advice and suggestions on how to make the review text better, mostly based on NLP models. However, whereas previous systems (discussed and named in Section 2.3) have implemented one of our two designs of text completion interfaces [17, 27, 55, 67, 93, 108, 115], the comparison between the two designs (*inline* and *list of suggestions*), especially regarding the task of peer review writing, is rare from an HCI perspective.

Hence, our research provides three contributions to the literature on intelligent user interfaces. First, by comparing two text completion designs used in previous works, we provided novel findings on how to embed NLP text generation models in user-centric assistants for writing peer reviews. Second, we collected user requirements from interviews with users having experience in writing peer reviews, and provided them to be used in future similar works on peer review writing assistants. Third, to evaluate our findings and to answer our research question, based on the collected user requirements, we designed a novel peer review writing assistant, `Hamta`, and measured how the users perceive it in our study. We believe designing peer review writing assistants based on our collected user requirements, as well as the findings in our study regarding the two conventional designs, offers a high potential to be accepted among reviewers in professional review writing settings.

6.2 Limitations and Future Work

Even so, our study has some shortcomings that present opportunities for future research. To begin with, the corpus we used for developing our tool [107] specifically included *business* peer reviews. In designing our system and in the user interviews as well as when searching for previous research, we focused on the domain-independent aspects of the peer review support tool, so that our provided design ideas would also be relevant for users who want to get assistance on writing scientific or online peer reviews. In addition, the tool we designed to find the answer to our research question (Hamta) can be easily transformed into a scientific or online peer review writing assistant only by changing the corpus used to fine-tune the GPT-2 model in its back-end. As a result, we think the design findings of this research are also applicable to researchers working in other domains of peer reviews. Nevertheless, we encourage researchers to rigorously investigate how the design findings from our research are relevant in the other peer review writing domains as well.

Moreover, we used a corpus of peer reviews in German collected in a university scenario by students [107], which is considered a very specific and limited setting, as all the students were German and belonged to the same university. Thus, future work can find corpora of peer reviews in other languages as well and add it to the data used to fine-tune our GPT-2 model, as in its current version, our tool, Hamta, can only provide generated suggestions in German. Therefore, using it can be inappropriate or confusing when evaluated by a different group of users from a different culture. Another point to consider is the relatively low number of participants in our experiment as well as the short duration of the experiment, which may not reflect the wide range of users of a peer review writing assistant in real environments accurately. Future studies should focus on how the results of this research can be replicated in different contexts and domains, as well as with more participants in a longer-term study.

The lowest value for the mean of the 1 to 7 Likert scale constructs we asked from the participants belonged to the mean of group B (*list of suggestions* design) on the construct *correctness of the suggestions* (M = 3.9, SD = 1.7)[26]. This relatively low value indicates a desire from the users for further improvement of the accuracy and relevance of the suggestions produced by our text generation model. This finding matches well with the previous research on ML-based systems, which claims users may not accept the suggestions and insights provided by imperfect AI systems and tools [59]. A possible remedy for this issue is to use newer, more performant models, such as GPT-3 [15][27], to improve the quality of the suggestions generated by our back-end text generation model as much as possible. Additionally, our study lacks measures to compare *review quality*, and rather chooses to only measure perception metrics, as well as the number of words as a quantitative indicator. We strongly encourage future works to also measure the quality of the outcome reviews, as well as a more diverse set of questions in the pre- and post-survey questionnaires from other validated sources. Lastly, we only compared and contrasted two designs of text completion interfaces, *inline* and *list of suggestions*, and we only conducted a between-subject evaluation, in which no participant tried both of the systems, due to our budget limits. Thus, we invite future researchers to consider other possible interfaces of text completion systems as well, as well as within-subject comparisons.

7 CONCLUSION

In our research work, we investigated the effects of embedding text generation models in user-centric assistants for writing peer reviews. To do so, we first explored the previous works on text completion systems and chose two designs to investigate further: *inline suggestions* versus providing a *list of suggestions*. To evaluate the similarities and differences between the two designs and find how users perceive them, we implemented them in Hamta, our peer review writing assistant with text generation models. We used a corpus from the literature [107] to fine-tune our GPT-2 model, and collected user requirements for implementing our user-centric generative peer review writing tool. We implemented our tool based on the requirements and evaluated both designs in a user study by asking the participants to provide a peer review on a business pitch. We found that the participants accepted our technology generally well and found it useful, exciting, and leading to improvements in their review's structure. Moreover, we found that while the participants who used the *inline suggestions* design provided a significantly higher number of words in their review text, those who used the *list of suggestions* design found the system significantly easier to use. The results from our study show that peer review writing assistants based on our findings can be used to assist users in writing peer reviews in professional contexts by generating and presenting example sentences. The users can directly insert the generated sentences into their reviews or alternatively take ideas from them. Additionally, we offer a set of user requirements we collected from our interviews with target users to be also used in designing future peer review writing assistants as well as user-centric systems with text generation models. Our results suggest including writing assistants based on our findings in collaborative learning environments and peer review processes.

REFERENCES

[1] Enrique Acosta, juan jose escribano otero, and Gabriela Toletti. 2014. Peer Review Experiences for MOOC . Development and Testing of a Peer Review System for a Massive Online Course. *The New Educational Review* 37 (10 2014), 66. https://doi.org/10.15804/tner.14.37.3.05

[2] Ritu Agarwal and Elena Karahanna. 2000. Time Flies When You're Having Fun: Cognitive Absorption and Beliefs about Information Technology Usage. *MIS Quarterly* 24, 4 (12 2000), 665. https://doi.org/10.2307/3250951

[3] Bruce Alberts, Brooks Hanson, and Katrina L. Kelner. 2008. Reviewing Peer Review. *Science* 321, 5885 (2008), 15–15. https://doi.org/10.1126/science.1162115 arXiv:https://www.science.org/doi/pdf/10.1126/science.1162115

[4] Beverly Alimo-Metcalfe. 1998. 360 Degree Feedback and Leadership Development. *International Journal of Selection and Assessment* 6, 1 (1998), 35–44. https://doi.org/10.1111/1468-2389.00070

[5] sharareh alipour, Sina Elahimanesh, Soroush Jahanzad, Parimehr Morassafar, and Seyed Parsa Neshaei. 2022. A Blockchain Approach to Academic Assessment. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 306, 6 pages. https://doi.org/10.1145/3491101.3519682

[26] While the mean for this construct in group A (*inline suggestions* design) was higher than the baseline (M = 4.2, SD = 1.8), the difference with group B is not statistically significant, which is in line with our expectations, as both designs used the same back-end with the same GPT-2 model.

[27] In our study, we did not have the necessary budget to also utilize GPT-3 APIs by OpenAI, so we call for future work to also evaluate GPT-3 in the task of peer review writing support and to explore the difference in the perceived correctness of the suggestions with the GPT-2 model among the users.

[6] Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 4699–4708.

[7] Yanti Andriyani, Rashina Hoda, and Robert Amor. 2017. Reflection in agile retrospectives. In *Agile Processes in Software Engineering and Extreme Programming (Lecture Notes in Business Information Processing)*, Hubert Baumeister, Horst Lichter, and Matthias Riebisch (Eds.). Springer, 3–19. https://doi.org/10.1007/978-3-319-57633-6_1 Conference on Agile Software Development 2017, XP 2017 ; Conference date: 22-05-2017 Through 26-05-2017.

[8] Kenneth C Arnold, Krzysztof Z Gajos, and Adam T Kalai. 2016. On suggesting phrases vs. predicting words for mobile text composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 603–608.

[9] Kathleen M Arnold, Sharda Umanath, Kara Thio, Walter B Reilly, Mark A McDaniel, and Elizabeth J Marsh. 2017. Understanding the cognitive processes involved in writing to learn. *Journal of Experimental Psychology: Applied* 23, 2 (2017), 115.

[10] S. J. Ashford. 1986. Feedback-Seeking in Individual Adaptation : A Resource Perspective. *Academy of Management Journal* 29, 3 (9 1986), 465–487. https://doi.org/10.2307/256219

[11] Kent Beck, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, and Dave Thomas. 2001. Manifesto for Agile Software Development. http://www.agilemanifesto.org/

[12] Xiaojun Bi, Tom Ouyang, and Shumin Zhai. 2014. Both complete and correct? Multi-objective optimization of touchscreen keyboard. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2297–2306.

[13] Baidyanath Biswas, Pooja Sengupta, Ajay Kumar, Dursun Delen, and Shivam Gupta. 2022. A critical assessment of consumer reviews: A hybrid NLP-based methodology. *Decision Support Systems* 159 (2022), 113799. https://doi.org/10.1016/j.dss.2022.113799

[14] James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. 2016. Quasi-Recurrent Neural Networks. https://doi.org/10.48550/ARXIV.1611.01576

[15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[16] William S Burroughs. 1961. The cut-up method of Brion Gysin. *The third mind* (1961), 29–33.

[17] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail Smart Compose. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. https://doi.org/10.1145/3292500.3330723

[18] Kun-Hung Cheng, Jyh-Chong Liang, and Chin-Chung Tsai. 2015. Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *The Internet and Higher Education* 25 (2015), 78–84. https://doi.org/10.1016/j.iheduc.2015.02.001

[19] Peng Cheng, Xiang Lian, Zhao Chen, Rui Fu, Lei Chen, Jinsong Han, and Jizhong Zhao. 2014. Reliable diversity-based spatial crowdsourcing by moving workers. *arXiv preprint arXiv:1412.0223* (2014).

[20] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *International conference on machine learning*. PMLR, 2067–2075.

[21] 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* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[22] Victoria Clarke and Virginia Braun. 2021. Thematic analysis: a practical guide. *Thematic Analysis* (2021), 1–100.

[23] Alan Cooper, Robert Reimann, and David Cronin. 2007. *About face 3: the essentials of interaction design*. John Wiley & Sons.

[24] Microsoft Corporation. 2023. *IntelliSense in Visual Studio Code*. https://code.visualstudio.com/docs/editor/intellisense

[25] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. 2018. Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine* 35, 1 (1 2018), 53–65. https://doi.org/10.1109/MSP.2017.2765202

[26] Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (1951), 297–334. https://doi.org/10.1007/BF02310555

[27] Arghavan Moradi Dakhel, Vahid Majdinasab, Amin Nikanjam, Foutse Khomh, Michel C Desmarais, Zhen Ming, et al. 2022. GitHub Copilot AI pair programmer: Asset or Liability? *arXiv preprint arXiv:2206.15331* (2022).

[28] Robert Dale. 2021. GPT-3: What's it good for? *Natural Language Engineering* 27, 1 (2021), 113–118.

[29] Richard Lee Davis, Thiemo Wambsganss, Wei Jiang, Kevin Gonyop Kim, Tanja Käser, and Pierre Dillenbourg. 2023. Fashioning the Future: Unlocking the Creative Potential of Deep Generative Models for Design Space Exploration. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–9.

[30] Gustavo H de Rosa and Joao P Papa. 2021. A survey on text generation using generative adversarial networks. *Pattern Recognition* 119 (2021), 108098.

[31] Mark D Dunlop and Andrew Crossan. 2000. Predictive text entry methods for mobile phones. *Personal Technologies* 4, 2 (2000), 134–143.

[32] Peter Elbow. 1998. *Writing with power: Techniques for mastering the writing process*. Oxford University Press.

[33] Katherine Elkins and Jon Chun. 2020. Can GPT-3 pass a Writer's turing test? *Journal of Cultural Analytics* 5, 2 (2020), 17212.

[34] Le Fang, Tao Zeng, Chaochun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. 2021. Transformer-based conditional variational autoencoder for controllable story generation. *arXiv preprint arXiv:2101.00828* (2021).

[35] Donald B Fedor, Kenneth L Bettenhausen, and Walter Davis. 1999. Peer reviews: Employees' dual roles as raters and recipients. *Group & Organization Management* 24, 1 (1999), 92–120.

[36] Tira Nur Fitria. 2021. Grammarly as AI-powered English writing assistant: Students' alternative for writing English. *Metathesis: Journal of English Language, Literature, and Teaching* 5, 1 (2021), 65–78.

[37] Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication* 32, 4 (1981), 365–387.

[38] Ana Frankenberg-Garcia, Robert Lew, Jonathan C Roberts, Geraint Paul Rees, and Nirwan Sharma. 2019. Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL* 31, 1 (2019), 23–39.

[39] Hansjörg Fromm, Thiemo Wambsganss, and Matthias Söllner. 2019. Towards A Taxonomy of Text Mining Features. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*.

[40] David Galbraith. 2009. Cognitive models of writing. *German as a foreign language* 2-3 (2009), 7–22.

[41] Hao Ge, Yin Xia, Xu Chen, Randall Berry, and Ying Wu. 2018. Fictitious GAN: Training GANs with historical models. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 119–134.

[42] Katy Gero, Alex Calderwood, Charlotte Li, and Lydia Chilton. 2022. A Design Space for Writing Support Tools Using a Cognitive Process Model of Writing. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. Association for Computational Linguistics, Dublin, Ireland, 11–24. https://doi.org/10.18653/v1/2022.in2writing-1.2

[43] Surjya Ghosh, Kaustubh Hiware, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. Does emotion influence the use of auto-suggest during smartphone typing?. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 144–149.

[44] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. https://doi.org/10.48550/ARXIV.1406.2661

[45] J Hattie and H Timperley. 2007. The power of feedback. Review of Educational Research, 77 (1), 81-112. Retrieved from. (2007).

[46] John R Hayes and Linda S Flower. 1980. The dynamics of composing: Making plans and juggling constraints. *Cognitive processes in writing* (1980), 31–50.

[47] John R Hayes and Linda S Flower. 1986. Writing research and the writer. *American psychologist* 41, 10 (1986), 1106.

[48] Yijue How and Min-Yen Kan. 2005. Optimizing predictive text entry for short message service on mobile phones. In *Proceedings of HCII*, Vol. 5. 2005.

[49] Julie S. Hui, Darren Gergle, and Elizabeth M. Gerber. 2018. IntroAssist: A Tool to Support Writing Introductory Help Requests. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173596

[50] Theresia Devi Indriasari, Andrew Luxton-Reilly, and Paul Denny. 2020. A Review of Peer Code Review in Higher Education. *ACM Trans. Comput. Educ.* 20, 3, Article 22 (sep 2020), 25 pages. https://doi.org/10.1145/3403935

[51] Touseef Iqbal and Shaima Qureshi. 2020. The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences* (2020).

[52] Ayush Jaiswal, Wael AbdAlmageed, Yue Wu, and Premkumar Natarajan. 2018. Capsulegan: Generative adversarial capsule network. In *Proceedings of the European conference on computer vision (ECCV) workshops*. 0–0.

[53] Tom Jefferson, Philip Alderson, Elizabeth Wager, and Frank Davidoff. 2002. Effects of editorial peer review: a systematic review. *Jama* 287, 21 (2002), 2784–2786.

[54] Jenni. 2023. *Jenni AI: Supercharge your writing with the most advanced AI writing assistant.* https://jenni.ai/

[55] Emil Thorstensen Jensen, Martin Hansen, Evelyn Eika, and Frode Eika Sandnes. 2020. Country selection on web forms: a comparison of dropdown menus, radio buttons and text field with autocomplete. In *2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. IEEE, 1–4.

[56] Jeff Johnson, Teresa L. Roberts, William Verplank, David Canfield Smith, Charles H. Irby, Marian Beard, and Kevin Mackey. 1989. The Xerox Star: A Retrospective. *Computer* 22, 9 (1989), 11–26.

[57] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635* (2018).

[58] Urvashi Khandelwal, Kevin Clark, Dan Jurafsky, and Lukasz Kaiser. 2019. Sample efficient text summarization using a single pre-trained transformer. *arXiv preprint arXiv:1905.08836* (2019).

[59] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of AI systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[60] Surafel M Lakew, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. *arXiv preprint arXiv:1806.06957* (2018).

[61] Severin Landolt, Thiemo Wambsganß, and Matthias Söllner. 2021. A Taxonomy for Deep Learning in Natural Language Processing. https://doi.org/10.24251/HICSS.2021.129

[62] Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent claim generation by fine-tuning OpenAI GPT-2. *World Patent Information* 62 (2020), 101983.

[63] Florian Lehmann and Daniel Buschek. 2022. Examining Autocompletion as a Basic Concept for Interaction with Generative AI. *CoRR* abs/2201.06892 (2022). arXiv:2201.06892 https://arxiv.org/abs/2201.06892

[64] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).

[65] Ming Liu, Rafael A Calvo, and Vasile Rus. 2012. Hybrid Question Generation Approach for Critical Review Writing Support. In *Proceedings of the 20th International Conference on Computers in Education. Singapore: Asia-Pacific Society for Computers in Education.*

[66] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. 2021. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834* (2021).

[67] Lawrence C Loh. 2016. Autocomplete: Dr Google's "helpful" assistant? *Canadian Family Physician* 62, 8 (2016), 622–623.

[68] Kristi Lundstrom and Wendy Baker. 2009. To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing* 18, 1 (2009), 30–43. https://doi.org/10.1016/j.jslw.2008.06.002

[69] Charles A. MacArthur. 1999. Word Prediction for Students with Severe Spelling Problems. *Learning Disability Quarterly* 22, 3 (1999), 158–172. https://doi.org/10.2307/1511283 arXiv:https://doi.org/10.2307/1511283

[70] A. J. Meadows. 1998. *Communicating Research.* Academic Press, San Diego, CA.

[71] Bhaskar Mitra, Milad Shokouhi, Filip Radlinski, and Katja Hofmann. 2014. On user interactions with query auto-completion. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval.* 1055–1058.

[72] Rohan Modi, Kush Naik, Tarjni Vyas, Shivani Desai, and Sheshang Degadwala. 2021. E-mail autocomplete function using RNN Encoder-decoder sequence-to-sequence model. In *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 710–714.

[73] Raoul A Mulder, Jon M Pearce, and Chi Baik. 2014. Peer review in higher education: Student perceptions before and after participation. *Active Learning in Higher Education* 15, 2 (2014), 157–171.

[74] David Nicol. 2014. Guiding principles for peer review: unlocking learners' evaluative skills. *Advances and Innovations in University Assessment and Feedback* (2014), 197–224.

[75] OECD. 2018. The Future of Education and Skills - Education 2030. https://doi.org/2018-06-15

[76] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 19, 10 pages. https://doi.org/10.1145/3411763.3450391

[77] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163.

[78] Baolin Peng and Kaisheng Yao. 2015. Recurrent Neural Networks with External Memory for Language Understanding. https://doi.org/10.48550/ARXIV.1506.00195

[79] Mariia Petryk, Michael Rivera, Siddharth Bhattacharya, Liangfei Qiu, and Subodha Kumar. 2022. How Network Embeddedness Affects Real-Time Performance Feedback: An Empirical Investigation. *Information Systems Research* 33, 4 (2022), 1467–1489. https://doi.org/10.1287/isre.2022.1110 arXiv:https://doi.org/10.1287/isre.2022.1110

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

[81] 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.

[82] Lloyd J. Rieber. 2006. Using Peer Review to Improve Student Writing in Business Courses. *Journal of Education for Business* 81, 6 (2006), 322–326. https://doi.org/10.3200/JOEB.81.6.322-326 arXiv:https://doi.org/10.3200/JOEB.81.6.322-326

[83] Roman Rietsche, Daniel Frei, Emanuel Stöckli, and Matthias Söllner. 2019. Not All Reviews are Equal - a Literature Review on Online Review Helpfulness. In *Proceedings of the 27th European Conference on Information Systems (ECIS).*

[84] Michael Rivera, Liangfei Qiu, Subodha Kumar, and Tony Petrucci. 2021. Are Traditional Performance Reviews Outdated? An Empirical Analysis on Continuous, Real-Time Feedback in the Workplace. *Information Systems Research* 32, 2 (2021), 517–540. https://doi.org/10.1287/isre.2020.0979 arXiv:https://doi.org/10.1287/isre.2020.0979

[85] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI EA '10)*. Association for Computing Machinery, New York, NY, USA, 2863–2872. https://doi.org/10.1145/1753846.1753873

[86] Taihua Shao, Yupu Guo, Honghui Chen, and Zepeng Hao. 2019. Transformer-based neural network for answer selection in question answering. *IEEE Access* 7 (2019), 26146–26156.

[87] Mike Sharples. 2022. Automated Essay Writing: An AIED Opinion. *International Journal of Artificial Intelligence in Education* (2022), 1–8.

[88] Xiaotian Su, Thiemo Wambsganss, Roman Rietsche, Seyed Parsa Neshaei, and Tanja Käser. 2023. Reviewriter: AI-generated instructions for peer review writing. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. 57–71.

[89] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification?. In *China national conference on Chinese computational linguistics*. Springer, 194–206.

[90] Vadim Sushko, Jurgen Gall, and Anna Khoreva. 2021. One-shot GAN: Learning to generate samples from single images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2596–2600.

[91] Alexey Svyatkovskiy, Shao Kun Deng, Shengyu Fu, and Neel Sundaresan. 2020. Intellicode compose: Code generation using transformer. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1433–1443.

[92] Jonas Thiergart, Stefan Huber, and Thomas Übellacker. 2021. Understanding Emails and Drafting Responses–An Approach Using GPT-3. *arXiv preprint arXiv:2102.03062* (2021).

[93] Tharis Thimthong, Thippaya Chintakovid, and Soradech Krootjohn. 2012. An empirical study of search box and autocomplete design patterns in online bookstore. In *2012 IEEE Symposium on Humanities, Science and Engineering Research*. 1165–1170. https://doi.org/10.1109/SHUSER.2012.6268796

[94] Almira Osmanovic Thunström and Steinn Steingrimsson. 2022. Can GPT-3 write an academic paper on itself, with minimal human input? (2022).

[95] Keith Topping. 1998. Peer Assessment Between Students in Colleges and Universities. *Review of Educational Research* 68, 3 (1998), 249–276.

[96] Keith J Topping. 2010. Methodological quandaries in studying process and outcomes in peer assessment. *Learning and instruction* 20, 4 (2010), 339–343.

[97] Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers.* " O'Reilly Media, Inc.".

[98] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *In Advances in neural information processing systems* Nips (2017), 5998–6008.

[99] Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences* 39, 2 (5 2008), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

[100] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly* 27, 3 (2003), 425–478.

[101] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[102] Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Jan Marco Leimeister, and Siegfried Handschuh. 2020. AL : An Adaptive Learning Support System for Argumentation Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing

Machinery, New York, NY, USA, 1–14.

[103] Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. A Corpus for Argumentative Writing Support in German. In *28th International Conference on Computational Linguistics (Coling)*. Barcelona, Spain. https://doi.org/10.18653/v1/2020.coling-main.74

[104] Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting Cognitive and Emotional Empathic Writing of Students. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 4063–4077. https://doi.org/10.18653/v1/2021.acl-long.314

[105] Thiemo Wambsganss, Matthias Soellner, Kenneth R Koedinger, and Jan Marco Leimeister. 2022. Adaptive Empathy Learning Support in Peer Review Scenarios. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 227, 17 pages. https://doi.org/10.1145/3491102.3517740

[106] Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Parsa Neshaei, Roman Rietsche, and Tanja Käser. 2023. Unraveling Downstream Gender Bias from Large Language Models: A Study on AI Educational Writing Assistance. *arXiv preprint arXiv:2311.03311* (2023).

[107] Thiemo Wambsganss, Vinitra Swamy, Roman Rietsche, and Tanja Käser. 2022. Bias at a Second Glance: A Deep Dive into Bias for German Educational Peer-Review Data Modeling. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, 1344–1356. https://aclanthology.org/2022.coling-1.115

[108] David Ward, Jim Hahn, and Kirsten Feist. 2012. Autocomplete as Research Tool: A Study on Providing Search Suggestions. *Information Technology and Libraries* 31, 4 (Dec. 2012), 6–19. https://doi.org/10.6017/ital.v31i4.1930

[109] Florian Weber, Thiemo Wambsganss, Seyed Parsa Neshaei, and Matthias Soellner. 2023. Structured persuasive writing support in legal education: A model and tool for German legal case solutions. In *Findings of the Association for Computational Linguistics: ACL 2023*. 2296–2313.

[110] World Economic Forum WEF. 2018. *The Future of Jobs Report 2018*. Technical Report. https://doi.org/10.1177/0891242417690604

[111] Vanessa Williamson. 2016. On the ethics of crowdsourced research. *PS: Political Science & Politics* 49, 1 (2016), 77–81.

[112] Matthew M. Yalch, Erika M. Vitale, and J. Kevin Ford. 2019. Benefits of Peer Review on Students' Writing. *Psychology Learning & Teaching* 18, 3 (2019), 317–325. https://doi.org/10.1177/1475725719835070 arXiv:https://doi.org/10.1177/1475725719835070

[113] Yu-Fen Yang. 2011. A reciprocal peer review system to support college students' writing. *British Journal of Educational Technology* 42, 4 (2011), 687–700. https://doi.org/10.1111/j.1467-8535.2010.01059.x arXiv:https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8535.2010.01059.x

[114] Su-Fang Yeh, Meng-Hsin Wu, Tze-Yu Chen, Yen-Chun Lin, XiJing Chang, You-Hsuan Chiang, and Yung-Ju Chang. 2022. How to Guide Task-Oriented Chatbot Users, and When: A Mixed-Methods Study of Combinations of Chatbot Guidance Types and Timings. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 488, 16 pages. https://doi.org/10.1145/3491102.3501941

[115] Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, and Zengfu Wang. 2009. Visual Query Suggestion. In *Proceedings of the 17th ACM International Conference on Multimedia* (Beijing, China) *(MM '09)*. Association for Computing Machinery, New York, NY, USA, 15–24. https://doi.org/10.1145/1631272.1631278

[116] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. 2015. Conditional Random Fields as Recurrent Neural Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 1529–1537. https://doi.org/10.1109/ICCV.2015.179

APPENDIX

Post-survey Questions

In Table 2, we present the constructs we used in the post-survey of our experiment, as well as the questions belonging to each of the constructs.

Table 2: Example questions for each of the constructs in our post-survey (translated to English and adapted to preserve their meaning in the context), as well as mean of the responses for each of our two groups (as presented in Section 4). We also used duplicate questions (with slight differences in wording in the German version) sparingly, which are indicated by (2x) in the table and averaged for reporting.

Construct	Example Questions	Mean A	Mean B
Technology Acceptance	Assuming `Hamta` is available, the next time I want to write a review, I would use it again. (2x)	4.11	4.21
Perceived Usefulness	With `Hamta` I can write reviews more effectively.	3.94	4.62
	I find using `Hamta` useful for writing reviews.	4.11	4.85
Perceived Review Quality	Compared to other participants, I think I wrote a very convincing review.	4.67	4.92
	I'm sure I wrote a very convincing review/ feedback.	4.72	4.69
	I'm sure I wrote a very insightful review/ feedback. (2x)	4.45	4.89
Perceived Ease of Use	Learning to use `Hamta` was easy for me.	4.44	5.15
	I find `Hamta` easy to interact with.	4.39	5.46
	It was easy for me to become familiar with `Hamta`.	4.94	5.46
Perceived Improvement in Writing Reviews	After using `Hamta`, my ability to write reviews has improved.	3.83	4.46
	After using `Hamta`, my ability to pay attention to the different parts of the review has improved.	4.11	4.54
Perceived Improvement in Reviews in the Long Run	I expect `Hamta` will help me improve my ability to write well-structured reviews.	4.28	4.92
	I expect `Hamta` will help me improve my ability to write helpful reviews.	4.17	4.62
	I assume `Hamta` would help me improve my ability to write compelling reviews.	4.22	4.46
	I assume `Hamta` would help me improve my ability to write insightful reviews.	4.11	4.54
Correctness of the Suggestions	`Hamta`'s suggestions are correct. (2x)	4.09	3.89
	The suggestions I received from `Hamta` were related to my text or my ideas.	4.33	3.92
Excitement After Interaction	Interacting with `Hamta` was fun and enjoyable for me.	4.50	4.38
	Interacting with `Hamta` was exciting.	4.17	4.31