Reviewriter: AI-Generated Instructions For Peer Review Writing

Xiaotian Su[1], Thiemo Wambsganss[1], Roman Rietsche [2], Seyed Parsa Neshaei[1], Tanja Käser[1]
[1] EPFL, Lausanne, Switzerland
{xiaotian.su, thiemo.wambsganss, seyed.neshaei, tanja.kaeser}@epfl.ch
[2] Universtiy of St.Gallen, St.Gallen, Switzerland
roman.rietsche@hsg.ch

Abstract

Large Language Models (LLMs) offer novel opportunities for educational applications that have the potential to transform traditional learning for students. Despite AI-enhanced applications having the potential to provide personalized learning experiences, more studies are needed on the design of generative AI systems and evidence for using them in real educational settings. In this paper, we design, implement and evaluate `Reviewriter`, a novel tool to provide students with AI-generated instructions for writing peer reviews in German. Our study identifies three key aspects: a) we provide insights into student needs when writing peer reviews with generative models which we then use to develop a novel system to provide adaptive instructions b) we fine-tune three German language models on a selected corpus of 11,925 student-written peer review texts in German and choose German-GPT2 based on quantitative measures and human evaluation, and c) we evaluate our tool with fourteen students, revealing positive technology acceptance based on quantitative measures. Additionally, the qualitative feedback presents the benefits and limitations of generative AI in peer review writing.

1 Introduction

Peer reviewing is a process by which learners provide formative feedback to each other on an individual task based on assessment criteria (Sadler and Good, 2006; Rietsche and Söllner, 2019). Research has found theoretical and empirical evidence for the positive effects of peer reviews on critical thinking skills (Lin et al., 2021; Ibarra-Sáiz et al., 2020), communication skills (Lai, 2016), and learning motivations (Hsia et al., 2016). The prevailing practice of peer review in tertiary education is evident in the eruption of massive open online courses (MOOCs) (Li et al., 2016). In these large-scale learning scenarios, peer review is particularly important since it is challenging for teachers to give effective one-by-one feedback due to immersive workload and shortage of time (Er et al., 2021). However, according to Oliver (1982), a challenge that plagues many student writers, including those having satisfactory grammar and spelling skills, is writer's block. It was defined by Rose (1980) as "that frustrating, self-defeating inability to generate the next line, the right phrase, the sentence that will release the flow of words again." A collaborator who provides instructions and points out new directions might help alleviate writer's block (Clark et al., 2018) and the combination of a writer's own ideas with suggested ideas is a form of psychological creativity (Boden et al., 2004). Novel LLMs have the potential to address the challenge of writer's block by generating suggestions for the next lines, right phrases, or sentences, thereby facilitating the flow of ideas (Gero et al., 2022), and helping students compose responses more efficiently (van Dis et al., 2023; Gao and Jiang, 2021). There are LLM-based collaborative writing tools to provide support for various writing tasks, including story writing (Yang et al., 2022), science writing (Gero et al., 2022), and screenwriting (Mirowski et al., 2022). However, few have investigated the utilization of generative AI for peer review writing tasks. Therefore, in this paper, we build and evaluate `Reviewriter` which can provide AI-generated instructions tailored to students' needs while writing peer reviews. It suggests possible directions based on students' input to inspire divergent outcomes while still leaving learners in control of the final text.

To investigate how to provide students with help to overcome writer's block in peer review writing, we conduct a literature review to gather insights for a peer review support system. We summarize five user requirements from interviews with twelve graduate students. Based on those, we develop seven design principles for providing AI-generated instructions in peer review tasks. Next, we search peer review corpora satisfying certain criteria and pre-process 11,925 student-written peer

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 57–71

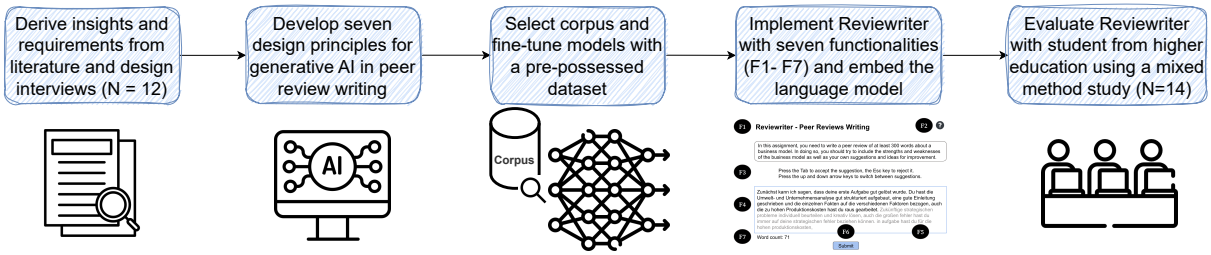

Figure 1: Overview of our methodology: We first gather system needs and requirements from literature and student interviews. Then we derive seven design principles with pedagogical considerations for a tool to provide AI-generated instructions for peer review writing tasks. Next, we fine-tuned three language models based on a selected corpus (Wambsganss et al., 2022b). Then, we instantiate the design in `Reviewriter` and evaluate it with fourteen students to assess its performance and gather quantitative as well as qualitative feedback.

review texts in German (Wambsganss et al., 2022b). We use it to fine-tune three language models to provide students with informative instructions. The best results according to training loss and human evaluation of fluency and correctness are achieved by German GPT-2. Then, we implement the design principles into the system to provide AI-generated instructions for peer review writing. Finally, in a mixed-method study with our full-working prototype, we evaluate the performance of the tool in a real-world learning exercise with fourteen students, and four of them also participated in the design interview. We assess the technology acceptance and level of enjoyment of the tool using well-defined constructs from Venkatesh and Bala (2008); Venkatesh et al. (2003) and also collect qualitative feedback from students.

Our research makes three contributions to the innovative use of NLP in education. Firstly, we provide insights and practical design considerations for incorporating AI-generated instructions in peer review writing tasks to overcome the known challenge of writer's block (Oliver, 1982). Secondly, we present and compare three open-source language models fine-tuned on a selected corpus of 11,925 student-written peer review texts in German. Lastly, we build `Reviewriter`, which implements seven functionalities with pedagogical design considerations and evaluates it on fourteen students from tertiary education. Our findings suggest that the tool providing AI-generated instructions in students' peer writing tasks leads to high ease of use and a high intention to use for students in their review writing process. Moreover, in the qualitative feedback, we find that the model has the potential to provide novel ideas for students to continue in depth. However, like other LLMs, it suffers from hallucination (Maynez et al., 2020) by producing factually incorrect and nonsensical answers, this invites further research to overcome and mitigate artificial hallucination. With `Reviewriter`, we present an interface with design rationales and an evaluated tool that other researchers can build upon to explore the effects of LLMs and the benefits and limitations of generative AI for writing peer reviews and building educational applications.

2 Related work

2.1 Student peer reviewing

There has always been significant interest in the study of peer reviews in the NLP community. Jia et al. (2022) introduced an approach called incremental zero-shot learning (IZSL) to address the issue of insufficient historical data for peer reviews. Wambsganss et al. (2022a) used empathy detection algorithms from NLP to analyze the given text and provide adaptive feedback in students' peer writing process. Moreover, several works have investigated how to embed classification models to support students in peer review writing. For example, researchers have explored the use of these models to develop argumentation skills (Wambsganss et al., 2020), support cognitive and emotional empathy writing (Wambsganss et al., 2021), and assess the specificity of written peer feedback (Rietsche et al., 2022). While NLP models, particularly LLMs, have the potential to deliver adaptive learning content (Adiguzel et al., 2023; Qadir, 2022), little research has focused on how to leverage their ability to provide tailored instructions for students during peer review writing (Darvishi et al., 2022). van Dis et al. (2023) mentioned benefits provided by generative AI for completing peer review tasks quickly. Experimental results from Gao and

Jiang (2021) showed that the effectiveness of generated suggestions, regardless of their performance quality, has consistently helped humans compose responses more efficiently when providing suggestions. In addition, Gero et al. (2022) demonstrated that students find it faster and easier to draw on language from generated texts than to write a sentence from scratch, even when given well-known information. Therefore, we propose a novel peer review writing tool `Reviewriter`, by leveraging the power of generative models, it can provide students with adaptive instructions to help them overcome writer's block in peer review writing.

2.2 NLP for writing support

With the massive success of ChatGPT, NLP is rapidly evolving as a key tool in writing support. On one hand, there is widespread adoption of generative AI in practice. Commercial writing assistants like Monica [1], a ChatGPT-powered extension, can support copywriting. And specialized applications like Jenni AI [2], Jasper AI [3] and Notion AI [4] can support creative writing. They are not only able to complete sentences but also generate the whole blog post and many other types of content including essays, emails, stories, and speeches based on users' input. On the other hand, many studies have focused on the use of language models for writing support in tertiary education. For instance, researchers have explored the use of these models for academic writing (Gero et al., 2022), fiction writing (Yang et al., 2022), and text summarization (Dang et al., 2022). Despite the widespread adoption of NLP in writing instruction, many models, including ChatGPT, remain general-purpose tools that have not been fine-tuned for specific tasks (Chen et al., 2023) or designed for particular educational settings (Kuhail et al., 2023). Embedding the AI techniques in a student-centered design is a complex task with several socio-technical challenges (Xu et al., 2021), including data collection (Zawacki-Richter et al., 2019), potential bias (Adiguzel et al., 2023) or discrimination (Pedróf et al., 2019) in the data, inadequate dataset training (Kuhail et al., 2023), incorporating the models, lack of student involvement in the design process (Verleger and Pembridge, 2018), lacking feedback on the generative system (Kuhail et al., 2023), and evaluating student perceptions (Xu et al., 2021). The present work provides insights into how to embed generative AI into peer review writing by establishing student-centered design with pedagogical considerations. We carefully select an unbiased corpus with a sufficient amount of peer review text to fine-tune language models. Furthermore, we evaluate student perceptions quantitatively and collect qualitative feedback on the generative AI system.

[1] https://monica.im/
[2] https://jenni.ai/
[3] https://www.jasper.ai
[4] https://www.notion.so/product/ai

3 Generative modeling to provide students adaptive instructions

3.1 The peer review dataset

To make sure our system is skilled in providing adaptive instructions for writing peer reviews and to improve accuracy and efficiency for human-AI interaction (Lee et al., 2022b), we decide to fine-tune language models with a peer review dataset. We start by searching the literature for a corpus that fulfilled the following criteria: a) it contains a large amount of student-written text in one particular domain (e.g., business model feedback) (Kuhail et al., 2023), b) it consists of a sufficient size to represent different nuances of characteristics in a balanced fashion (e.g. specificity, helpfulness) (Rietsche et al., 2022), and c) it does not possess a significant bias (e.g. gender, racial or social discrimination) (Adiguzel et al., 2023). The business model peer review corpus published in Wambsganss et al. (2022b) fulfilled all these requirements. The corpus consists of 11,925 peer reviews collected at a university in the German-speaking area of Europe. They were written by first-year master's students in a business department course. The student population has an average age of 24.6 years old with a standard deviation of 1.7 years. Students wrote approximately 9 peer reviews per course with an average length of 220 words. Furthermore, Wambsganss et al. (2022b) showed that this collected corpus does not reveal many biases in nine WEAT co-occurrence analyses or in the GloVe embeddings. This corpus provides us with a sufficient amount of unbiased peer review texts to fine-tune language models for adaptive instructions in the domain of business peer reviews.

3.2 Data pre-processing

To ensure the model could generate high-quality instructional text, we select reviews written from 2016 to 2021 with a rated helpfulness score greater than five on a 1 - 7 Likert Scale (1: low, 4: neutral,

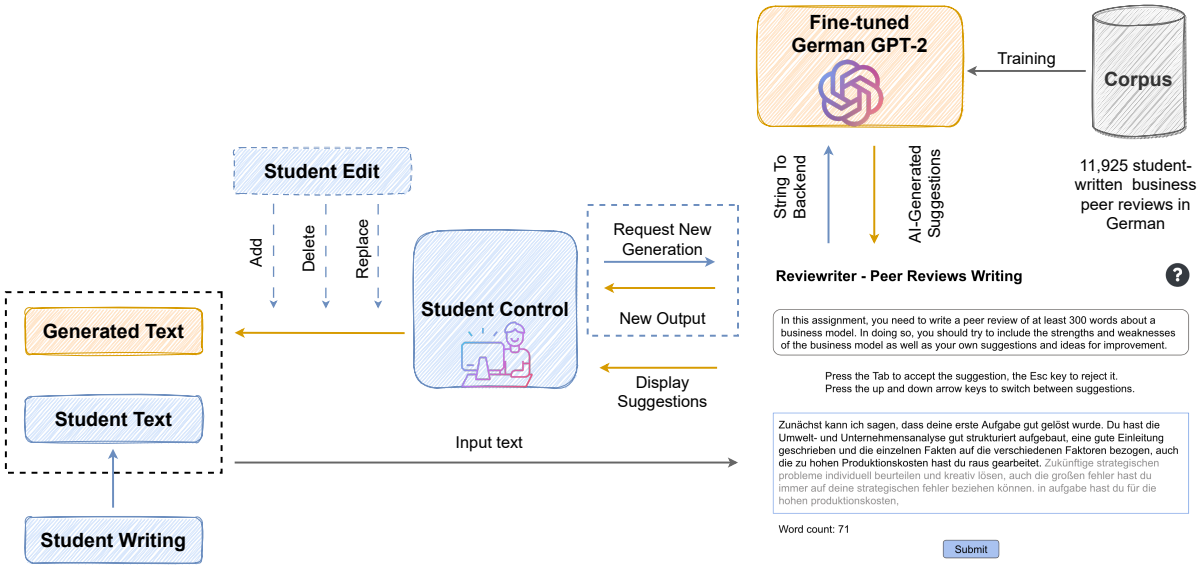

Figure 2: Architecture of `Reviewriter` to provide AI-generated instructions for students to write peer reviews. First, students enter initial input, which is then used by the German GPT-2 model to generate instructions. The students evaluate the generated content and decide whether to regenerate it. Following this, students are free to edit the instructions. Finally, both the generated text and the student's text are utilized as inputs for the next generation.

7: high). We start by removing HTML tags, irrelevant information like PDF file names and specific information like URLs, keywords (revealing the identity of students), and questions asked to write reviews which some students copied to their review text (Appendix A.1). We also expand abbreviations as shown in Appendix A.2. Then, we shuffle and divide cleaned data into train and test datasets with proportions of 0.8 and 0.2 for fine-tuning and evaluating the language model. Lastly, all sentences are tokenized with model-specific tokenizers.

3.3 The generative models

Transformer-based language models, such as BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019), using the pre-training and fine-tuning paradigm, have revolutionized NLP and achieved state-of-the-art records on various tasks. These models are first pre-trained in a self-supervised fashion on a large corpus and fine-tuned for specific downstream tasks (Wang et al., 2018). In our case, to provide AI-generated instructions for German peer review writing, we use pre-trained causal language models on the HuggingFace platform (Wolf et al., 2020) for German text generation. We choose them because there is no usage limitation and by utilizing open-source technology, we contribute to LLM transparency (van Dis et al., 2023; Adiguzel et al., 2023), allowing other researchers to easily replicate our findings or build upon them. Therefore, we selected two German GPT-2 models (`dbmdz/german-gpt2` [5] and `benjamin/gerpt2-large` [6]) and one multilingual model BLOOM (Scao et al., 2022) (`bigscience/bloom-560m` [7]). We did not use GPT3 for fine-tuning since it was not open-source available at the time of our research. For all of them, we fine-tune the pre-trained models following the default hyperparameter settings (Appendix A.3) with block size 128, and 500 warm-up steps.

We compare training loss and used human evaluation to select the best model. Note that GerPT2-large already performs well (Appendix A.4 for sample generated text) after ten epochs of training, even with higher training loss compared to the other two models (Table 1). However, it suffers a long inference time (a student needs to wait around 10 seconds to get instructions given 40 words) compared to the other two models (5 seconds with the same input). Therefore, we decide to further evaluate German GPT-2 and BLOOM. We conduct a human evaluation of the quality of the generated response. Specifically, we sample ten instructions generated by each model and present them to two German researchers to evaluate their fluency and correctness. From the evaluation of both parties, German GPT-2 yields more coherent results than

[5] https://huggingface.co/dbmdz/german-gpt2
[6] https://huggingface.co/bigscience/bloom-560m
[7] https://huggingface.co/benjamin/gerpt2-large

the BLOOM model and there are more meaningless sentences from the response generated by BLOOM than by German GPT-2. Therefore, we decide to use the German GPT-2 model as the base for the tool with a default temperature of 1.0 for generating the next token.

PLM	Size # Param.	Training loss	Training epochs
German GPT-2	124	0.0418	30
BLOOM	560M	0.0560	30
GerPT2-large	774M	2.8183	10

Table 1: Comparison of the number of parameters for three transformer-based pretrained language models (PLMs) and their training and evaluation loss.

3.4 The generative system

To design a system providing AI-generated instructions for peer review writing, we first draw on insights from relevant literature. Following the methodology of Cooper (1988), we analyze human-AI interaction (Shen and Wu, 2023; Chan et al., 2023; Lee et al., 2022b) and NLP-supported peer review systems (Alqassab et al., 2023; Darvishi et al., 2022). Then, to gather insights into the needs of writing peer reviews with AI-generated instructions for tertiary education, we conduct semi-structured interviews with twelve graduate students. We reach out to a group of computer science students who previously registered in a business class and have experience writing peer reviews on business models, and to students in our university for general recruitment. The participants have a diverse background in computer science, business, or psychology, and a mean age of 24.50 years (SD = 2.02), including two females and ten males (representing the distribution of computer science students at our school). Half of them had experience writing peer reviews, while the others did not. Each interview lasts around 30 to 50 minutes. We use the expert qualitative interview method outlined in Brinkmann (2013) and Gläser and Laudel (2009) to gain an initial understanding of students' needs for receiving adaptive instructions in peer review writing. We ask topics about prior experience with technology-based writing systems, perceptions of existing writing systems (e.g., Grammarly), difficulties in writing peer reviews, and desired functionalities for a system to support peer review writing. We transcribe the interviews and identify five clusters of requirements following Cohn (2004). We find that 75% of the students would like to interact with a clean and straightforward interface (*user requirement - UR 1*). Two-thirds of interviewees asked for intuitive guidance on how to interact with the tool (*UR 2*). And 41.7% of them said that they would like to see more than one instruction to choose from (*UR 3*). One-third of the students stated that they prefer to view a complete piece of instruction rather than words or phrases to formulate a concrete idea (*UR 4*). Lastly, two-thirds of them indicated that they would like to see the number of words they have entered to have better control over the structure of the review (*UR 5*).

	Design Principle
DP1)	Provide a web-based application with a responsive clean and intuitive interface to allow students to use the tool with ease and stay motivated to write.
DP2)	Provide clear and detailed guidance to ensure that students understand how to use the tool and can take full advantage of the features offered.
DP3)	Provide an intuitive keyboard control to make it easy for students to manipulate the AI-generated instructions.
DP4)	Provide a simple text area for students to write, edit the peer review, and view multiple inline instructions.
DP5)	Present instructions in an inline format in the text area to help students quickly pick up ideas while allowing them to stay in the context of writing to reduce cognitive burden.
DP6)	Provide a complete argument for each instruction to assist students in constructing comprehensive reviews.
DP7)	Present a summary of statistics on the text to guide students on how many words they have written.

Table 2: Derived design principles on how to provide AI-generated instructions for students to write peer reviews.

With insights derived from the literature review and requirements from student interviews (similar to Rietsche et al. (2018)), we develop seven design principles (Table 2) and further map them to seven functionalities (Figure 3 *F1 - F7*) in `Reviewriter`, a responsive web application to provide AI-generated instructions for peer review writ-

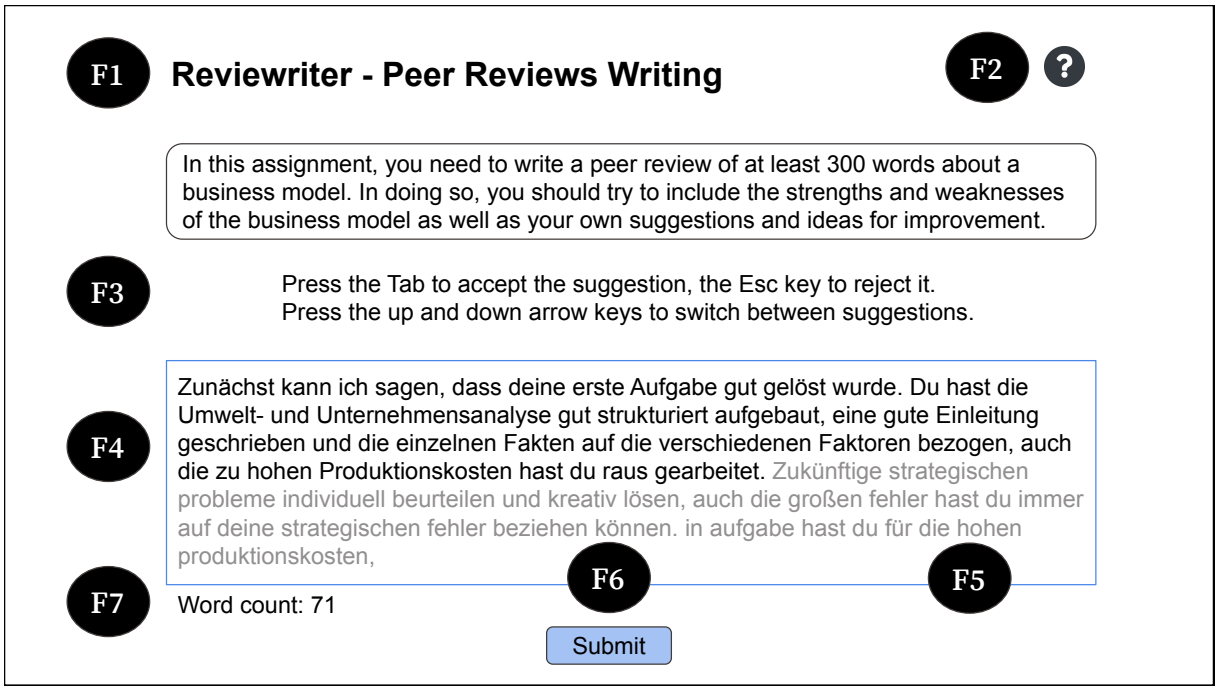

Figure 3: A screenshot of `Reviewriter` and its main functionalities (*F1 - F7*) derived from system requirements and design principles. The system provides a clean interface (*F1*). By clicking the question mark, students get detailed guidance on the peer review writing task and the usage of the tool (*F2*). A simple text area supports all typical interactions, such as typing, selecting, editing, and deleting text, and caret movement via keys and mouse (*F4*). In the input area, the sentences in black are the actual text, we display the AI-generated instruction in an inline format in gray (*F5*). The model generates next-sentence predictions to give students a complete view of the idea (*F6*). We provide three instructions each time, and students may use the *Tab* key to accept, the *Esc* key to reject, and the *Up* and *Down* arrow keys to toggle through different instructions (*F3*). The total number of words is displayed below the text area to inform students of their writing progress (*F7*).

ing. The design is student-centered and has two main components: a neat interface with key commands for text editing (Figure 3) and a generative language model in the backend 3.3. To foster the independent thinking of students and discourage overreliance on technology (Adiguzel et al., 2023), we organize a workshop with two senior researchers to deliberate on the optimal timing for presenting the generated instructions. Combined with studies Buschek et al. (2021); Bhat et al. (2021), we decide to present instructions until students have entered a minimum number of words and put a certain amount of delay before showing instructions to minimize potential disruptions caused by irrelevant information from model hallucination (Maynez et al., 2020). Figure 2 presents the system architecture. The student starts with writing the beginning of the review. The system will display instructions until students enter at least 25 words. After this threshold, when the student gets stalled, by pressing the spacebar, they will trigger the model in the backend to generate instructions. After the keypress, there is a delay of eight seconds before they receive instructions. To preserve the context while avoiding too much overhead for querying the mode, we pass the last twenty words from the input to the model. According to *UR 4*, and supported by Calderwood et al. (2020), overly brief suggestions are often unhelpful. To ensure clarity and concision, we limit each instruction to a maximum of 60 tokens, which is approximately 45 words [8]. In their experiment with one, three, and six instructions, Buschek et al. (2021) discovered that multiple instructions can facilitate the identification of useful phrases and boost their acceptance rate. We decide to present three instructions each time considering the cost-benefit tradeoffs for efficiency (e.g. reading time vs diversified content). The student controls the final output by checking multiple instructions and deciding whether to accept or reject them. They are free to add, delete, and replace the generated content.

[8] https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

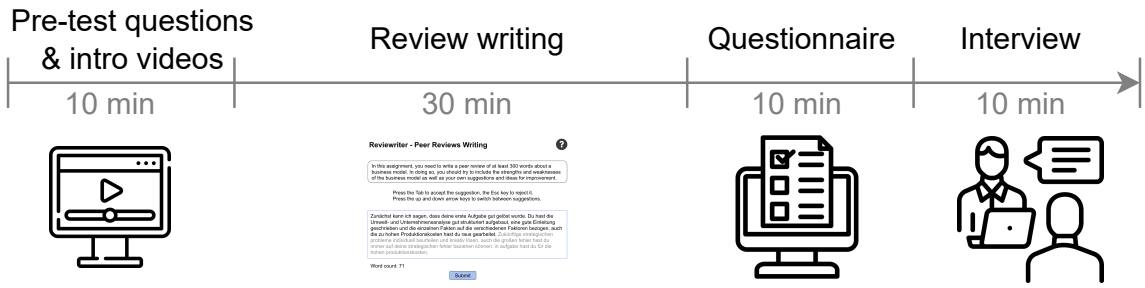

Figure 4: Overview of the study procedure. Students begin with five pre-test questions and two introduction videos. Then, they engage in a 30-minute review writing task. Afterward, they are asked to complete a questionnaire, which is followed by an interview with a set of open-ended questions.

4 Evaluation of Reviewriter

4.1 Experimental setup

To assess our prototype, we conduct a mixed-method study with fourteen students from a public university in Europe. We reached out to students who have participated in our previous design interview and also recruited students on campus. Fourteen students–eleven males and three females–participated in the evaluation. Three of them were undergraduate students and the rest were graduate students. Four graduate students also participated in our previous design interview. They were all native German speakers and expressed interest in getting AI-generated instructions when writing peer reviews. They have diverse backgrounds, including computer science, robotics, and business with a mean age of 25.33 years (SD = 3.60). The evaluation is conducted either face-to-face or remotely with a conference tool. Each student screen records their writing process, the interviews are also recorded and transcribed by a researcher.

1. Pre-test (10 minutes): The experiment starts with a pre-survey that has five questions (Appendix B.1) followed by two videos. The first four questions measure the learners' level of innovation in the field of information technology, following Agarwal and Karahanna (2000). They need to rate their agreement with a statement on a Likert scale ranging from 1 (totally disagree) to 7 (totally agree), with 4 being neutral (Likert, 1932). Following the pre-survey, we present two videos. The first video introduces a business model for a platform that connects ski instructors with learners, and the second video provides guidance on how to use `Reviewriter`.

2. Peer review writing (30 minutes): In this phase, students are asked to write a review for a peer's business model. Specifically, they are asked to elaborate on strengths, weaknesses, and suggestions for improvement of the given business model. We instruct students not to use search engines and spend a minimum of 15 minutes on the task. A countdown indicates the remaining time.

3. Questionnaire and interview (10+10 minutes): In the post-survey, we ask 29 questions (Appendix B.2) to measure *perceived ease of use*, *perceived ease of interaction*, *perceived level of enjoyment*, *perceived level of excitement* and *perceived usefulness*, following the technology acceptance model of Venkatesh and Bala (2008) and Venkatesh et al. (2003). All constructs are measured with a 1- to 7-point Likert scale. Moreover, we ask several qualitative questions to further examine students' attitudes toward AI-generated instructions and capture the demographics.

4.2 Quantitative analysis and qualitative feedback

To measure student perceptions of AI-generated instructions for peer review writing, we calculate the following constructs on a 1- to 7-point Likert scale (Table 3): perceived ease of use ($M_1 = 6.07, SD_1 = 0.83$), perceived ease of interaction ($M_2 = 5.50, SD_2 = 1.22$), perceived level of excitement ($M_3 = 5.64, SD_3 = 1.15$), perceived level of enjoyment ($M_4 = 5.43, SD_4 = 1.16$), and perceived usefulness ($M_5 = 4.64, SD_5 = 1.34$). The results show that the participants rate positively using `Reviewriter` to receive adaptive instructions. Moreover, the mean values of the tool are also very promising when comparing the results

Statistics	**Perceived ease of use**	**Perceived ease of interaction**	**Perceived level of excitement**	**Perceived level of enjoyment**	**Perceived usefulness**
Mean	6.07	5.50	5.64	5.43	4.64
Std.	0.83	1.22	1.15	1.16	1.34
Normalized mean	0.87	0.79	0.81	0.78	0.66

Table 3: Descriptive statistics from quantitative measure in the evaluation of `Reviewriter` (N=14). The measure of technology acceptance on a 1 - 7 Likert Scale (1: low, 4: neutral, 7: high).

to the average of the scale. All results are better than the neutral value of four. This fosters motivation and engagement to use the learning application. Malik et al. (2021) found that perceived ease of use ($M_1 = 6.07$) and usefulness ($M_5 = 4.64$) positively influence student adoption intentions and their attitudes toward AI-based applications. The positive levels of perceived ease of interaction ($M_2 = 5.50$), excitement ($M_3 = 5.64$), and enjoyment ($M_4 = 5.43$) suggest that the technology has been accepted favorably. This is especially important for learning tools to ensure students are perceiving the usage of the tool as enjoyable, useful, and easy to interact with (Marangunić and Granić, 2015). These are promising results for using this tool to receive AI-generated instructions in a peer review setting.

In addition to quantitative scores, we incorporate qualitative open-ended questions to further understand student attitudes toward writing with AI-generated text and how the instructions impact their writing process. We translate the responses from German and cluster the representative ones (Appendix B.3). The general attitude towards `Reviewriter` was very positive. Five students stated concretely the benefits of `Reviewriter` on their writing process. Three students mentioned the system is simple and easy to interact with. On the adoption of the generated instructions, one student used them every time, two students stated that they did not find anything useful in the instructions. Another two students reported that they never used the complete instructions but they picked up ideas or keywords from them. Five of them used instructions three to five times, and the rest stated that they use the AI-generated instructions quite frequently and did not provide an exact number. Moreover, it is interesting to note that there are divergent opinions on the delay of the system. Three students complained about the waiting time was too long while two other students were in favor of the delay and stated that the waiting time left them room to think. Finally, students enjoyed the diverse content in AI-generated instructions while noticing there were ungrammatical sentences and irrelevant phrases from time to time.

5 Discussion

Peer review writing is an increasingly important educational task in large-scale or distance learning scenarios since it enables personalized feedback to be delivered at scale, thereby lessening the workload of instructors (Er et al., 2021) and boosting learners' motivation (Hsia et al., 2016). However, during writing peer reviews, students may experience obstacles such as writer's block Rose (1980) where they struggle to generate the next line, the right phrase, or the sentence Oliver (1982). LLMs can help to overcome this obstacle by producing adaptive instructions based on students' input, which ultimately aid in the seamless progression of thoughts (Gero et al., 2022). To do so, we develop a novel peer review writing tool called `Reviewriter`. It allows students to use AI-generated instructions as an inspiration and incorporate those ideas into their own work in a creative and original way, such as by adapting, mixing, or reinterpreting those instructions (Qadir, 2022).

Our study contributes at least three key aspects to the innovative use of NLP in education. First, we explore the personalization of AI-generated instructions in a specific pedagogical scenario - peer review writing (Pardos and Bhandari, 2023) by gathering insights from literature review and student interviews (Verleger and Pembridge, 2018). Second, in contrast to Lee et al. (2022a) which used GPT-3 without adaptation for collaborative writing, we fine-tune three German language models on a corpus selected based on certain criteria to provide specialized content with high quality. Afterward, we choose German-GPT2 based on quantitative measures and human evaluation. Third, as noted

in Kuhail et al. (2023), "lack of feedback" is one of the challenges to using generative models in education. Therefore, we evaluate our tool with fourteen students and the result reveals positive technology acceptance based on quantitative measures. Through our qualitative evaluation, we find that students generally enjoyed seeing generated instructions with varied content to spark ideas. And they were enthusiastic and excited about writing with generative language models. We recognize that there is a need for further research on the effectiveness of LLM-based writing support tools in various contexts, as well as the improvement of faithfulness and factuality in AI-generated instructions (Maynez et al., 2020). Nonetheless, our study contributes to the growing body of knowledge on the potential of generative AI to provide personalized writing instructions and enhance students' learning experiences (Pardos and Bhandari, 2023).

6 Conclusion and future work

To help students mitigate writer's block during peer review writing, we design, build, and evaluate `Reviewriter`, a novel tool that aims to provide students with AI-generated instructions during their peer review writing process. We provide design insights with pedagogical considerations of integrating LLMs into peer-review writing systems. Our evaluation involves fourteen students from tertiary education, who reported enjoying the interaction with the system, finding it easy to use, and expressing interest in using similar tools in the future. They also pointed out that the relevance of the generated instructions could be further improved. We present `Reviewriter`, including its design rationales and evaluated interface, as a contribution to the exploration of LLMs' potential in innovative NLP-based approaches in education. As NLP continues to advance, we aspire that our work will encourage other researchers to explore how generative AI can be integrated into educational applications to benefit teachers and students, while promoting responsible and ethical use.

For future work, we will investigate students' perceptions of peer reviews from different sources: their peers, peers using `Reviewriter`, and entirely AI-generated reviews. We will collect ratings and feedback from students who receive these reviews and compare the relevance, quality, and usefulness of the texts generated from each source. Additionally, we aim to integrate `Reviewriter` into the university's existing peer review system, enabling widespread adoption among students across various courses. By incorporating AI-generated instructions into routine peer reviews, we can examine the long-term impact on student's writing skills, critical thinking abilities, and overall academic performance. To enhance the relevance of the AI-generated instructions in `Reviewriter`, we will refine the algorithms and models based on feedback from our evaluation participants. Our iterative development process will involve incorporating more contextual information, employing advanced NLP techniques, and leveraging user feedback to achieve higher accuracy and helpfulness in the AI-generated instructions.

References

Tufan Adiguzel, Mehmet Haldun Kaya, and Fatih Kürşat Cansu. 2023. Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology* 15, 3 (2023), ep429.

Ritu Agarwal and Elena Karahanna. 2000. Time flies when you're having fun: Cognitive absorption and beliefs about information technology usage. *MIS quarterly* (2000), 665–694.

Maryam Alqassab, Jan-Willem Strijbos, Ernesto Panadero, Javier Fernández Ruiz, Matthijs Warrens, and Jessica To. 2023. A systematic review of peer assessment design elements. *Educational Psychology Review* 35, 1 (2023), 18.

Advait Bhat, Saaket Agashe, and Anirudha Joshi. 2021. How do people interact with biased text prediction models while writing?. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Online, 116–121. https://aclanthology.org/2021.hcinlp-1.18

Margaret A Boden et al. 2004. *The creative mind: Myths and mechanisms*. Psychology Press.

Svend Brinkmann. 2013. *Qualitative interviewing*. Oxford university press.

Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages. https://doi.org/10.1145/3411764.3445372

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

Zijian Ding Chan et al. 2023. Mapping the Design Space of Interactions in Human-AI Text Co-creation Tasks. *arXiv preprint arXiv:2303.06430* (2023).

Yu Chen, Scott Jensen, Leslie J Albert, Sambhav Gupta, and Terri Lee. 2023. Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers* 25, 1 (2023), 161–182.

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

Mike Cohn. 2004. *User stories applied: For agile software development*. Addison-Wesley Professional.

Harris M Cooper. 1988. Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in society* 1, 1 (1988), 104–126.

Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 98, 13 pages. https://doi.org/10.1145/3526113.3545672

Ali Darvishi, Hassan Khosravi, Solmaz Abdi, Shazia Sadiq, and Dragan Gašević. 2022. Incorporating Training, Self-Monitoring and AI-Assistance to Improve Peer Feedback Quality *(L@S '22)*. Association for Computing Machinery, New York, NY, USA, 35–47. https://doi.org/10.1145/3491140.3528265

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Erkan Er, Yannis Dimitriadis, and Dragan Gašević. 2021. Collaborative peer feedback and learning analytics: Theory-oriented design for supporting classwide interventions. *Assessment & Evaluation in Higher Education* 46, 2 (2021), 169–190.

Zihan Gao and Jiepu Jiang. 2021. Evaluating Human-AI Hybrid Conversational Systems with Chatbot Message Suggestions. In *Proceedings of the 30th ACM International Conference on Information amp; Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 534–544. https://doi.org/10.1145/3459637.3482340

Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing Using Language Models. In *Designing Interactive Systems Conference* (Virtual Event, Australia) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1002–1019. https://doi.org/10.1145/3532106.3533533

Jochen Gläser and Grit Laudel. 2009. *Expert interviews and qualitative content analysis: as tools for reconstructive research*. Springer-Verlag.

Lu-Ho Hsia, Iwen Huang, and Gwo-Jen Hwang. 2016. Effects of Different Online Peer-Feedback Approaches on Students' Performance Skills, Motivation and Self-Efficacy in a Dance Course. *Comput. Educ.* 96, C (may 2016), 55–71. https://doi.org/10.1016/j.compedu.2016.02.004

María Soledad Ibarra-Sáiz, Gregorio Rodríguez-Gómez, and David Boud. 2020. Developing student competence through peer assessment: the role of feedback, self-regulation and evaluative judgement. *Higher Education* 80, 1 (2020), 137–156.

Qinjin Jia, Yupeng Cao, and Edward Gehringer. 2022. Starting from "Zero": An Incremental Zero-shot Learning Approach for Assessing Peer Feedback Comments. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. Association for Computational Linguistics, Seattle, Washington, 46–50. https://doi.org/10.18653/v1/2022.bea-1.8

Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Education and Information Technologies* 28, 1 (2023), 973–1018.

Chin-Yuan Lai. 2016. Training nursing students' communication skills with online video peer assessment. *Computers & Education* 97 (2016), 21–30.

Mina Lee, Percy Liang, and Qian Yang. 2022a. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. 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 388, 19 pages. https://doi.org/10.1145/3491102.3502030

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong,

et al. 2022b. Evaluating Human-Language Model Interaction. *arXiv preprint arXiv:2212.09746* (2022).

Hongli Li, Yao Xiong, Xiaojiao Zang, Mindy L. Kornhaber, Youngsun Lyu, Kyung Sun Chung, and Hoi K. Suen. 2016. Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education* 41, 2 (2016), 245–264.

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

Hui-Chen Lin, Gwo-Jen Hwang, Shao-Chen Chang, and Yaw-Don Hsu. 2021. Facilitating critical thinking in decision making-based professional training: An online interactive peer-review approach in a flipped learning context. *Computers & Education* 173 (2021), 104266.

Reena Malik, Ambuj Shrama, Sonal Trivedi, and Rikkee Mishra. 2021. Adoption of Chatbots for learning among university students: role of perceived convenience and enhanced performance. *International Journal of Emerging Technologies in Learning (iJET)* 16, 18 (2021), 200–212.

Nikola Marangunić and Andrina Granić. 2015. Technology acceptance model: a literature review from 1986 to 2013. *Universal access in the information society* 14, 1 (2015), 81–95.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2022. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals. *arXiv preprint arXiv:2209.14958* (2022).

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

Zachary A Pardos and Shreya Bhandari. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871* (2023).

Francesc Pedróf, Miguel Subosa, Axel Rivas, and Paula Valverde. 2019. Artificial intelligence in education : challenges and opportunities for sustainable development.

Junaid Qadir. 2022. Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. (2022).

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.

Roman Rietsche, Andrew Caines, Cornelius Schramm, Dominik Pfütze, and Paula Buttery. 2022. The Specificity and Helpfulness of Peer-to-Peer Feedback in Higher Education. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. Association for Computational Linguistics, Seattle, Washington, 107–117. https://doi.org/10.18653/v1/2022.bea-1.15

Roman Rietsche, Kevin Duss, Jan Martin Persch, and Matthias Soellner. 2018. Design and Evaluation of an IT-based Formative Feedback Tool to Foster Student Performance. In *Proceedings of the International Conference on Information Systems (ICIS)*. San Francisco, CA, USA.

Roman Rietsche and Matthias Söllner. 2019. Insights into Using IT-Based Peer Feedback to Practice the Students Providing Feedback Skill. Proceedings of the Hawaii International Conference on System Sciences (HICSS), Maui, HI, USA.

Mike Rose. 1980. Rigid Rules, Inflexible Plans, and the Stifling of Language: A Cognitivist Analysis of Writer's Block. *College Composition and Communication* 31, 4 (1980), 389–401. http://www.jstor.org/stable/356589

Philip M Sadler and Eddie Good. 2006. The impact of self-and peer-grading on student learning. *Educational assessment* 11, 1 (2006), 1–31.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100* (2022).

Hua Shen and Tongshuang Wu. 2023. Parachute: Evaluating Interactive Human-LM Co-writing Systems. *arXiv preprint arXiv:2303.06333* (2023).

Eva AM van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. 2023. ChatGPT: five priorities for research. *Nature* 614, 7947 (2023), 224–226.

Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences* 39, 2 (2008), 273–315.

Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.

Matthew Verleger and James Pembridge. 2018. A pilot study integrating an AI-driven chatbot in an introductory programming course. In *2018 IEEE frontiers in education conference (FIE)*. IEEE, 1–4.

Thiemo Wambsganss, Andrew Caines, and Paula Buttery. 2022a. ALEN App: Argumentative Writing Support To Foster English Language Learning. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. Association for Computational Linguistics, Seattle, Washington, 134–140. https://doi.org/10.18653/v1/2022.bea-1.18

Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. AL: An Adaptive Learning Support System for Argumentation Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376732

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting cognitive and emotional empathic writing of students. *arXiv preprint arXiv:2105.14815* (2021).

Thiemo Wambsganss, Vinitra Swamy, Roman Rietsche, and Tanja Käser. 2022b. 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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, 353–355. https://doi.org/10.18653/v1/W18-5446

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wei Xu, Marvin J Dainoff, Liezhong Ge, and Zaifeng Gao. 2021. From human-computer interaction to human-AI Interaction: new challenges and opportunities for enabling human-centered AI. *arXiv preprint arXiv:2105.05424* 5 (2021).

Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and Ray LC. 2022. AI as an Active Writer: Interaction strategies with generated text in human-AI collaborative fiction writing. In *Joint Proceedings of the ACM IUI Workshops*, Vol. 10.

Olaf Zawacki-Richter, Victoria Marín, Melissa Bond, and Franziska Gouverneur. 2019. Systematic review of research on artificial intelligence applications in higher education -where are the educators? *International Journal of Educational Technology in Higher Education* 16 (10 2019), 1–27. https://doi.org/10.1186/s41239-019-0171-0

A Details on data pre-processing and models

A.1 Template questions asked students to write reviews which some students copied to their review text

- What do you see as the strengths of the fellow student's solution?
- What do you see as weaknesses in the fellow student's solution and how can they be addressed?
- What should be paid attention to in the revision of the solution?
- Provide concrete suggestions for improvement in this regard.
- Give concrete suggestions for improvement (constructive feedback).
- What should you pay attention to in the revision of the solution? Give concrete suggestions for improvement (constructive feedback).

A.2 Abbreviations and expansions

Abbreviation	Expansion
bsp, bspw	beispielsweise
dh	da her
ev, evtl	eventuell
ggf	gegebenenfalls
oä	oder ähnliches
vlt	vielleicht
zb	zum Beispiel

Table 4: A list of abbreviations students used in the review text and we replace with the expansion in the pre-processing.

A.3 Hyperparameters for pretrained language models

Hyperparameter	GPT2	BLOOM
Vocabulary size	50257	250880
Attention heads	12	8
Hidden layers	12	2
Attention dropout	0.1	0.1

Table 5: Hyperparameters for pretrained GPT2 and BLOOM

A.4 Sample text generated by different language models

B Details on evaluations

B.1 Pre-test questions asked during evaluation of `Reviewriter`

1. I like experimenting and trying out new technologies.
2. As a rule, I am hesitant when trying out new technologies.
3. In my circle of friends, I'm usually the first person to try new digital media / new technologies.
4. When I hear about new technologies I look for a way to experiment with them.
5. I have had experience writing reviews/feedback in the past.

B.2 Post-test questions asked during evaluation of `Reviewriter`

- Transition questions: How many times have you accepted `Reviewriter`'s recommendations?
- Technology Acceptance Model
  1. Assuming the review writing assistance tool is available, the next time I want to write a review/feedback I would use it again.
  2. With `Reviewriter` I can write reviews/feedback more effectively.
  3. Learning to use `Reviewriter` was easy for me.
  4. I find using `Reviewriter` useful for writing reviews/feedbacks.
  5. I find `Reviewriter` easy to interact with.
  6. It would be easy for me to become familiar with `Reviewriter`.
  7. Compared to other participants, I think I wrote a very convincing review/feedback.
  8. After using `Reviewriter`, my ability to write reviews/feedback has improved.
  9. I'm sure I wrote a very insightful review/feedback.
  10. I'm sure I wrote a very convincing review/feedback.

11. With Reviewriter I can write better reviews/ feedbacks.
12. I think I now know more about how to write well-structured, persuasive, and insightful reviews/feedbacks.
13. Assuming Reviewriter was available, the next time I write a review/feedback I would use it.
14. After using Reviewriter, my ability to pay attention to the different parts of the review/feedback structure has improved.

- Evaluate student perceptions on the AI-generated instructions
  1. I expect Reviewriter will help me improve my ability to write well-structured reviews/feedbacks.
  2. I assume Reviewriter would help me improve my ability to write compelling reviews/feedback.
  3. I assume Reviewriter would help me improve my ability to write insightful reviews/feedback.
  4. Interacting with the tool was fun and enjoyable for me.
  5. I expect Reviewriter will help me improve my ability to write helpful reviews/feedback.
  6. Interacting with the tool was exciting.

- Open-ended questions for qualitative feedback
  1. How has Reviewriter impacted your writing process?
  2. What did you particularly like about using Reviewriter?
  3. Do you have any other ideas?
  4. What could still be improved?
  5. Have you used a writing support program before (e.g. Grammarly)?
  6. What is your field of study?
  7. Please enter your gender.
  8. Please indicate your mother tongue.

B.3 Clustered qualitative student feedback from the evaluation of Reviewriter

Source	Text
Original input	First of all, I can say that your first task was solved well. You have structured the environmental and company analysis well, written a good introduction, and related the individual facts to the various factors, and you have also worked out the excessively high production costs.
German GPT-2 generated instructions	Furthermore, the problems were identified and their elaboration was well coordinated. However, the behaviour of this company is completely unnecessary and one could hardly judge it more. Unfortunately, you went into far too much detail about the individual points and did little about the environmental and company analysis. 2) What do you see as the weaknesses of the environmental and business analysis and how can these be remedied?1.written out above:However, with the exception of case-related, these are other generic terms that have a rather different linguistic form.
BLOOM generated instructions	I have already recognized the second task and the reasoning is also understandable for you. Good luck for the future! Strengths) The project idea is cleanly and realistically implemented, as it is a
GerPT2-large generated instructions	You mastered the task very well. I have a few suggestions for improvement. First of all, you could have explained in the assignment what an environmental and company analysis actually is. I would also replace the word "company" with "firm" or something similar. You answered task two well, but you might have briefly explained the stakeholder approach, which was necessary to understand the task. Perhaps you could have changed the focus of the company analysis, but that is only a small point. In general you could

Table 6: Comparison of the sample text generated by different transformer-based pretrained language models with max_length=150 (we translated all text from German to English for the purpose of this paper).

Topic	Cluster	Statement
On the adoption of the AI-generated instructions	Positive	S1: "I mainly accepted the ideas and slightly rewrote the proposed text." S3: "I find myself be inspired by professional keywords." S11: "I used the recommendations every time."
	Constructive	S4: "Never. They were utterly useless."
On the quality of the AI-generated instructions	Positive	S1: "A few of the suggested ideas were very relevant. It also often remind me to say something positive." S4: " I like that it suggests diverse ideas that are quite different from each other." S10: "Reviewriter provided me with novel ideas that I could explore in depth."
	Constructive	S1: "Shorter instructions would be more relevant sometimes." S10: "The instructions sometimes have spelling mistakes." S11: " Sometimes I got instructions that didn't fit the content." S12: "I would suggest to generate shorter snippets. Sometimes the beginning wasn't bad but later it got weird."
On the impact of the writing process	Positive	S2: "The tool helps break through writer's block." S3: " When I got stuck on what to write, it sometimes had useful keywords, which made me a little quicker." S10: "The review writing process has accelerated." S11: "I got new ideas from Reviewirter's suggestions. I think the system not only helps to write structured reviews, but also to come up with new ideas. This is where I see the greatest potential." S14: " I didn't feel so alone while writing."
	Constructive	S1: "Waiting for suggestions slowed down my writing process." S12: "I tried to adopt the instructions a couple of times to be more efficient. However, since the waiting time for the instructions is very long, the process has been delayed."
On the system interaction	Positive	S5, S8: "It is easy to use and simple to operate." S10: "It is easy to use and saves time." S11: "I liked that I was not forced to accept the instructions and I could choose among several options."
	Constructive	S11: "I think it would be better if we could select the instructions with the mouse."
On the delay of instructions	Positive	S2: "Latency is moderate." S9: "I did not get suggestions instantaneously, I really just got it when I wanted it. That was really good, because that way my thoughts did not get interrupted." S14: "It is good that the instructions don't come immediately after I stop writing. It didn't disrupt my flow of writing."
	Constructive	S6: "The proposals come too late, I almost come up with my own ideas." S1, S10, S12: "The waiting time for suggestions is long."

Table 7: We have categorized the qualitative feedback received from fourteen students (referred to as S1 to S14) from tertiary education, who participated in the evaluation of `Reviewriter`. We collected the feedback through open-ended questions in the post-survey and concluding interview. For qualitative questions answered in German, we translated the written responses into English. The interview was conducted in English, recorded with the students' consent.