PEER: Empowering Writing with Large Language Models

Kathrin Seßler[0000−0002−3380−4641], Tao Xiang, Lukas Bogenrieder, and Enkelejda Kasneci[0000−0003−3146−4484]

Technical University of Munich, Munich, Germany
kathrin.sessler@tum.de

Abstract. The emerging research area of large language models (LLMs) has far-reaching implications for various aspects of our daily lives. In education, in particular, LLMs hold enormous potential for enabling personalized learning and equal opportunities for all students. In a traditional classroom environment, students often struggle to develop individual writing skills because the workload of the teachers limits their ability to provide detailed feedback on each student's essay. To bridge this gap, we have developed a tool called PEER (Paper Evaluation and Empowerment Resource) which exploits the power of LLMs and provides students with comprehensive and engaging feedback on their essays. Our goal is to motivate each student to enhance their writing skills through positive feedback and specific suggestions for improvement. Since its launch in February 2023, PEER has received high levels of interest and demand, resulting in more than 4000 essays uploaded to the platform to date. Moreover, there has been an overwhelming response from teachers who are interested in the project since it has the potential to alleviate their workload by making the task of grading essays less tedious. By collecting a real-world data set incorporating essays of students and feedback from teachers, we will be able to refine and enhance PEER through model fine-tuning in the next steps. Our goal is to leverage LLMs to enhance personalized learning, reduce teacher workload, and ensure that every student has an equal opportunity to excel in writing. The code is available at https://github.com/Kasneci-Lab/AI-assisted-writing.

Keywords: Large Language Models · Writing · Personalized Education

1 Introduction

The introduction of transformers-based technologies [13] for natural language processing (NLP) has been a breakthrough that pushed the field significantly forward. It enabled the development of pre-trained large language models (LLMs) which can process natural language more effectively and efficiently than previous approaches [1, 10]. The most recent models, like ChatGPT [8], have been fine-tuned using reinforcement learning with human feedback [9], enhancing their ability to generate human-like conversations and leading to a wide range of novel applications and use cases in various domains, also in the field of education [5].

Since LLMs are trained to write high quality texts, they can assist users in their writing process [15]. More specifically, LLM-based tools can help improve writing skills already from the very young age up to professional writing.

During their academic years, students are learning various types of essays. However, in the traditional classroom setting, teachers are not able to provide detailed feedback for each student's work due to time constraints and heavy workload. Also, feedback is usually only given once (e.g. in the context of graded homework or assessments) without further possibility to enhance the writing afterwards and receive an updated feedback, impeding a continuous process of improvement.

To tackle this challenge in essay writing education, and hence support both learners and teachers, we have developed an AI-based tutor named PEER, Paper Evaluation and Empowerment Resource. The idea behind PEER is to offer comprehensive textual feedback on the learner's essay, including specific suggestions for improvement, while being always constructive, specific and engaging. This stands in contrast to previous work, where the focus was often on merely grading the essay rather than offering comprehensive feedback [11]. PEER also allows students to make adjustments to their work and receive updated feedback to provide an ongoing process of improvement. From an educator's perspective, PEER provides an initial structure and suggestions for constructive and thorough feedback that can serve as a basis for further enhancements by the teacher. Such AI-assisted feedback can save a lot of time and energy, reducing hence the teacher's workload and offering more space for interaction with the students. Existing work that uses LLMs to improve writing skills often focuses more on general strategies and does not have a concrete focus on the different essay types that are part of the school curriculum with their demands and challenges [12].

The difficulty of this project lies in the limited availability of student's essay data and corresponding feedback of the teachers. Due to privacy reasons, such data is typically not publicly available in vast amounts, making it less likely for large language models to have encountered this type of data during their pre-training phase. However, the use of LLMs eliminates the need for costly hand-crafted features that previously formed the basis of many automated essay scoring systems [4]. PEER explores how the capabilities of LLMs can be leveraged to assist students and teachers in personalizing essay writing education, which involves teaching the model to provide reasonable and helpful feedback tailored to different types of essays.

We argue that by interacting with PEER, a learner can gain the necessary skills to comprehend the essential elements of good essay writing, and thereby enhance their own writing abilities. In order to evaluate the efficacy and implications of PEER, we have initiated collaborations with various German educators, schools, and academic institutions who expressed interest in our endeavor.

While the current version of PEER has been designed specifically to hone writing abilities in the German language and for the German educational system, we envision expanding its applicability to encompass other pedagogical frameworks and languages.

2 PEER

PEER is a user-friendly, web-based tool designed to analyze students' essays and generate comprehensive feedback that includes concrete suggestions and engaging tips to improve writing.

2.1 Pedagogical Background

Providing effective feedback is crucial for enhancing students' learning experience [3], as it can significantly impact their performance and motivation [7]. However, traditional classroom settings often do not allow teachers to provide comprehensive and engaging feedback to each student, hindering the learning process. To address these challenges, PEER adopts two main approaches:

1. **Thorough and Constructive Feedback**: PEER provides detailed and constructive feedback that is always positive, engaging, and helps to motivate students. The feedback highlights the strengths of the students' work and provides suggestions for improvement. PEER also uses a visual indicator (green instead of red color) to represent feedback that is encouraging and supportive.
2. **Continuous Learning Environment**: PEER follows a continuous learning principle that allows students to receive feedback as many times as they require without any fear of being discouraged or judged. This approach fosters a safe and barrier-free learning environment where students can learn at their own pace, and feedback is always readily available.

2.2 Technical Background

Our approach aims to provide students with comprehensive and constructive feedback on their essays. In the absence of a task-specific data set, we achieve this by combining zero-shot learning (inference without prior training only based on semantic information) with Elo ratings (numerical system to compare the relative performance) for different prompts. Through collecting feedback from the users on the quality of the output, we can continuously improve our model.

Zero-shot learning Traditionally, machine learning models have been trained to perform on specific tasks. One significant advantage of LLMs is their versatility [10], including zero-shot learning, which is the ability to follow textual instructions [1]. Finding the optimal instructions already evolved into an own field of research [14, 6] and several heuristics can aid in crafting suitable prompts. By providing the right instructions, LLMs can be guided to generate feedback that closely resembles that of a teacher. For example, a possible prompt can look like the following (translated from German into English):

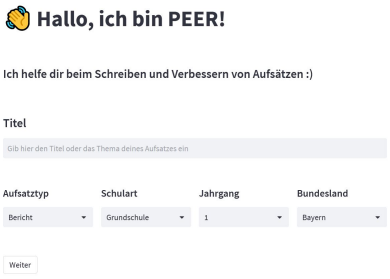
```
The following text is a {article_type} from a student in the
{year}th grade. {extra_infos} The topic is "{title}". Text:
"{essay}". Analyze the text based on the criteria and give
feedback and suggestions for improvement like a teacher.
```

Prompt Elo rating Our approach involves systematically identifying the optimal model instructions from a set of prompts through user feedback, utilizing the Elo rating system [2]. When a user requests feedback on an essay, two different responses generated from distinct instructional prompts are presented. Based on the user's preference, the ratings for the corresponding instructional prompts are updated using the Elo system. By incorporating human feedback into the process, our system continuously improves, guided by the success of reinforcement learning with human feedback deployed to fine-tune ChatGPT [9].

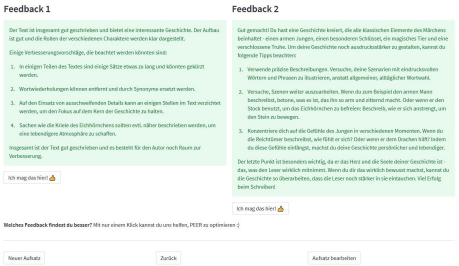
Weighted Lottery System When generating new feedback, the two instructional prompts that wrap the user's essay are selected using a weighted lottery system based on their respective Elo ratings. This ensures that prompts with higher ratings have a greater chance of being chosen from the set, resulting in feedback that is more likely to be of superior quality. At the same time, the use of a random selection process ensures that all prompts are evaluated.

3 Prototype

To make PEER available for all students, we have developed a website and are currently working on an accompanying application, in order to further reduce the barriers to access and ensure widespread availability.



(a) The Start Page. In the first step, the topic of the essay and the relevant meta data is entered.



(b) The Feedback Page. The user is provided with two feedback texts and can mark the preferred one to improve our model.

In the first step, users can input the topic or title of their essay, along with relevant meta information such as essay type, school year, and school type. Then, they can choose to insert the text manually or upload an image, which is scanned using an OCR and post-processed by GPT-3 [1] to remove any artifacts from the image. Next, PEER evaluates the input and generates two feedback texts. The users are then encouraged to indicate which feedback they find more useful. To facilitate continuous learning and improvement, the users can modify their essay according to the feedback and request new feedback. This process can be repeated as many times as necessary to enhance the writing skills of the users.

4 Preliminary Results

Over 4000 essays have already been uploaded for evaluation, with argumentation being the most frequently requested category. The platform was primarily used by students from middle and upper levels.

Prompt Evaluation Based on the Elo scores, prompting the model to approach the task as a friendly teacher and providing it with additional information about the specific essay type leads to the best results. It enables the model to focus on the relevant characteristics and use the extra information to improve and adapt its feedback accordingly.

Feedback from Teachers The quantitative results are complemented by feedback from teachers who assessed the tool from a qualitative perspective. Several teachers reported to us their experiences of trying PEER themselves as well as applying it together with their students in their classrooms. Overall, they acknowledged PEER's usefulness for both students and teachers, highlighting its user-friendliness, respectful tone, and timely feedback that facilitates individualized learning. However, they also identified some areas for improvement. For instance, they noted that the feedback provided by PEER can be too general at times, such as suggesting to "use more adjectives." Additionally, one teacher pointed out that in the German language, both the male and female forms for professions are typically used to be inclusive. Unfortunately, PEER currently does not account for both versions in its output text, and sometimes even marks them as redundant in students' essays. Other criticisms included missing essay types on the start page and sometimes small grammar errors in the generated feedback texts. This initial qualitative assessment and teacher feedback is incorporated in the further development of PEER before a more comprehensive and larger user study is conducted.

5 Conclusion and Future Agenda

Based on the amount of feedback we have received from teachers so far, it is clear that PEER is meeting a need in schools for both learners and teachers. However, our project is still in its early stages and requires further development, model fine-tuning and user evaluations. As our objective is to bring PEER to schools and establish it as a valuable assistant in the process of learning how to write an essay, our concrete current and next steps are as follows:

- Creating a solid data basis consisting of real-world essays and high-quality feedback provided by teachers. This data set will then be used to fine-tune a large language model for our specific domain to improve the performance.
- Conducting a user study at schools to assess the performance of PEER and gather valuable insights into the positive and negative aspects of the tool for both teachers and students.

Given the inherent stochastic nature of the underlying model, it is important to acknowledge that a fully error-free outcome cannot be guaranteed. Consequently, and as for all applications based on large language models, users are advised to carefully evaluate the feedback provided and selectively incorporate only the pertinent critiques. This type of critical thinking is not limited to PEER but should be a fundamental aspect of interacting with any generative AI model.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
2. Elo, A.E.: The rating of chessplayers, past and present. Arco Pub. (1978)
3. Hattie, J., Timperley, H.: The power of feedback. Review of educational research **77**(1), 81–112 (2007)
4. Hussein, M.A., Hassan, H., Nassef, M.: Automated language essay scoring systems: A literature review. PeerJ Computer Science **5**, e208 (2019)
5. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences **103**, 102274 (2023)
6. Liu, J., Shen, D., Zhang, Y., Dolan, W.B., Carin, L., Chen, W.: What Makes Good In-Context Examples for GPT-3? In: Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. pp. 100–114 (2022)
7. Molloy, E.K., Boud, D.: Feedback models for learning, teaching and performance. Handbook of research on educational communications and technology pp. 413–424 (2014)
8. OpenAI Team: ChatGPT: Optimizing language models for dialogue (2022)
9. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
10. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
11. Ramesh, D., Sanampudi, S.K.: An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review **55**(3), 2495–2527 (2022)
12. Schick, T., Dwivedi-Yu, J., Jiang, Z., Petroni, F., Lewis, P., Izacard, G., You, Q., Nalmpantis, C., Grave, E., Riedel, S.: PEER: A Collaborative Language Model. arXiv preprint arXiv:2208.11663 (2022)
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
14. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E.H., Le, Q.V., Zhou, D., et al.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: Advances in Neural Information Processing Systems (2022)
15. Yuan, A., Coenen, A., Reif, E., Ippolito, D.: Wordcraft: story writing with large language models. In: 27th International Conference on Intelligent User Interfaces. pp. 841–852 (2022)