Assessing Online Writing Feedback Resources: Generative AI vs. Good Samaritans

Shabnam Behzad[1], **Omid Kashefi**[2], **Swapna Somasundaran**[2]

[1] Georgetown University, USA
[2] Educational Testing Service (ETS), USA
shabnam@cs.georgetown.edu, {okashefi, ssomasundaran}@ets.org

Abstract

Providing constructive feedback on student essays is a critical factor in improving educational results; however, it presents notable difficulties and may demand substantial time investments, especially when aiming to deliver individualized and informative guidance. This study undertakes a comparative analysis of two readily available online resources for students seeking to hone their skills in essay writing for English proficiency tests: 1) essayforum.com, a widely used platform where students can submit their essays and receive feedback from volunteer educators at no cost, and 2) Large Language Models (LLMs) such as ChatGPT. By contrasting the feedback obtained from these two resources, we posit that they can mutually reinforce each other and are more helpful if employed in conjunction when seeking no-cost online assistance. The findings of this research shed light on the challenges of providing personalized feedback and highlight the potential of AI in advancing the field of automated essay evaluation.

Keywords: Computer-Assisted Language Learning, Usability, Evaluation Methodologies

1. Introduction

Providing feedback on students' essays is vital for their learning journey, as it fosters self-awareness, enhances their overall learning experience, and encourages iterative growth. Constructive feedback guides students toward specific enhancement strategies and reinforces the concept of continuous improvement. Additionally, feedback clarifies assignment expectations and cultivates essential skills such as critical thinking, self-assessment, and effective communication, which are invaluable for their future careers (McMillan, 1987; Farra et al., 2015; Cajander et al., 2015).

However, delivering such feedback is challenging due to the need for personalized, motivating, and constructive guidance while managing time and resources effectively. Ensuring timely and personalized feedback that aligns with students' unique learning styles and needs further compounds this challenge. A question that this study aims to address is to what extent Large Language Models (LLMs) can facilitate this process.

The development of LLMs has led to substantial advances in NLP, creating opportunities for educational technology (Caines et al., 2023; Hicke et al., 2023; Baffour et al., 2023; Xiao et al., 2023; Jeon and Lee, 2023; Huang et al., 2023). One such opportunity pertains to the generation of automatic essay feedback, which can be valuable for both educators and students. Guo and Wang (2023) investigated ChatGPT's performance in generating feedback on students' writing and compared it with professional teacher-generated feedback in terms of their length and type and the potential of GPT-generated feedback to support teachers in their feedback provision. Escalante et al. (2023) also suggest a blended approach that combines the strengths of both AI-generated and human tutor feedback could offer a promising solution.

The diverse array of LLMs applications necessitates different evaluation studies (Beigman Klebanov and Madnani, 2020). Our primary focus in this study is on students who, due to the cost-effectiveness and around-the-clock accessibility of online resources, increasingly seek writing feedback from such sources. In particular, we study and compare two of these resources, used by language learners preparing for English proficiency tests: (i) essayforum.com[1] which is a collaborative online community forum to receive and give feedback on submitted essays and (ii) ChatGPT, an LLM which is freely available online (Ouyang et al., 2022).

The study provided in this paper assesses and compares GPT-generated feedback and human feedback on five core aspects of crafting effective written responses: *relevance*, *highlighting strength*, *highlighting weakness*, *being specific*, and *overall helpfulness* (Ende, 1983; Ovando, 1994; Omer and Abdularhim, 2017). Our studies show that each resource has its advantages and disadvantages, with many instances highlighting their complementary nature. It is essential to underscore that the objective of this study is not to assess the potential of AI as a substitute for human educators. Instead, the aim is to investigate its prospective role in complementing and enhancing human feedback.

[1] https://essayforum.com/writing/

	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Task description	**0.819**	0.279	0.034	0.135	0.020
Rubric-based	0.814	**0.289**	**0.043**	**0.140**	**0.035**
Few-shot examples	**0.819**	0.246	0.029	0.138	0.013

Table 1: Automatic evaluation results (reference-based) for essay feedback generation task. Significant differences are indicated by numbers in bold, determined using the Mann–Whitney test.

	Coherence	Consistency	Fluency	**Overall**
Human feedback	0.796	0.506	0.791	0.763
Task description	**0.971**	0.651	**0.953**	**0.877**
Rubric-based	0.965	**0.702**	0.876	0.868
Few-shot examples	0.956	0.673	0.944	0.873

Table 2: UniEval results (referenceless evaluation) for essay feedback generation task

2. Experiments

In this section, we describe the methodology and results of our experiments designed to evaluate the effectiveness of GPT-generated feedback for English language learner essays, comparing different prompt paradigms and their alignment with human feedback.

2.1. Data

Since there is no publicly available data for English language learner essays and feedback comments, we collect our data from essayforum.com, a community forum that has sections for different types of writing. We specifically focus on the *writing feedback* forum since it's mainly used by students practicing for English proficiency tests. Previous work has used data from this site for different purposes. Stab et al. (2014) released a subset of essays from this site annotated with argument components and argumentative relations (Stab and Gurevych, 2014). Bao et al. (2022) also constructed a dataset from this site for argumentative essay writing task.

We adapted preprocessing steps from Bao et al. (2022) since their focus is the essay itself and not the feedback comments. The preprocessing steps included removing essays that needed a reference picture and some cleaning steps to remove noisy data. Then, we randomly selected 300 pairs of essays and feedback for our experiments and conducted a deeper study on 30% of them in human evaluation. Essays in this set have an average length of 432 words and feedback comments have an average length of 283 words.

2.2. Experimental Setup

We used GPT-3.5 in our experiments to mimic the performance of ChatGPT. We followed different prompt designing paradigms for generating feedback for a given student essay, including in-context learning and prompt-tuning (Brown et al., 2020; Liu et al., 2023). In this work, we report the results for these three types of prompting strategies:

Task description. Simply describing the feedback generation task:

```
You are an AI assistant that helps students
practicing for English proficiency tests improve
their essays. I will give you an essay written by a
student. Write constructive feedback explaining
how to improve the essay as a teacher would.
```

Rubric-based. Specifically mentioning aspects that are used in rubrics for grading:

```
You are an AI assistant that helps students practicing
for English proficiency tests improve their essays.
I will give you an essay written by a student.
Write constructive feedback, explaining how to
improve the essay like a teacher would do. You should
focus on the following aspects in your feedback:
1- Is the essay in response to the topic?
2- Are there spelling, grammar, or stylistic issues?
3- Is the point of view of the essay author clear?
4- Are there clear and descriptive arguments that
support the essay author's position or point of view?
```

Few-shot. Provide 3 example essays, including one strong, one weak, and an average essay, and their feedback at inference time.

2.3. Results

We report the results of these experiments in Table 1. Based on these results, overall, *rubric-based* prompt seems to be the most similar to human feedback available on essayforum.com. Since the aim of this work is to compare feedback from humans and GPT on different aspects, we also use UniEval (Zhong et al., 2022) *unieval-sum* pretrained model which evaluates coherence, consistency, and fluency of text without needing a gold human reference. The results are available in Table 2. Surprisingly, we observe that considering all aspects, GPT achieves a higher score than

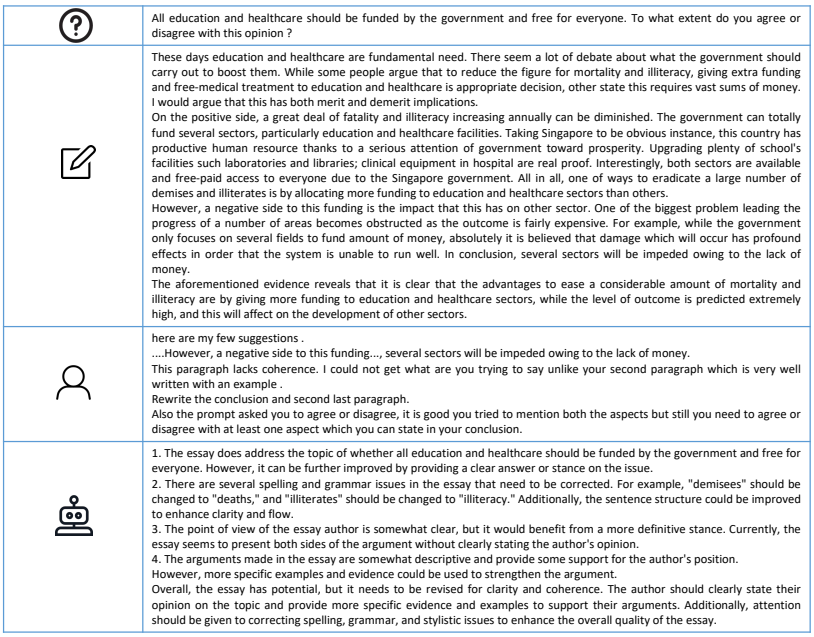

Figure 1: Example essay with two feedback comments: Human feedback and GPT-3.5 rubric-based

human-generated feedback. One possible reason could be that unieval-sum model was not specially trained for this task, but more importantly, the results from Tables 1 and 2 confirm the well-known limitations of automatic metrics for evaluating generation tasks (Xu et al., 2023; Kasai et al., 2022; Bhakthavatsalam et al., 2021; Celikyilmaz et al., 2020). Due to the variability observed across different metrics, we believe that conducting a human evaluation for deeper analysis is necessary for this task.

3. Human Evaluation

Human evaluation was done by two annotators (one male and one female), who have experience in both NLP research and the educational domain. Annotators were also familiar with English proficiency tests as they are non-native English speakers (high level of English proficiency) who both have previous experience with taking these tests. Multiple training rounds and discussions were done until annotators reached an agreement above 65% in all categories (Cohen's Kappa) which we believe is reasonable for a task that could be subjective in many cases.

During the training phase, annotators were each assigned a set of 10 essays with four corresponding feedback, one from humans and three from LLM with different prompting strategies. They were asked to rank feedback comments based on five criteria:

- **C1**: Which feedback is more **relevant** to the essay content?
- **C2**: Which feedback is better at highlighting **weakness**?
- **C3**: Which feedback is better at highlighting **strengths**?
- **C4**: Which feedback is more **specific** and **actionable**?
- **C5**: Which feedback is overall more **helpful** for a student?

These criteria are synthesized from the literature on *constructive feedback* (Ende, 1983; Ovando, 1994; Omer and Abdularhim, 2017), except for the first one which we specifically added to see if there are any hallucinations in GPT-generated feedback. The average time for reading one essay, all feedback comments, and ranking was 15 minutes.

For the final round of annotations, we chose the best GPT feedback based on training phase results which was the feedback from the *rubric-based* prompt. In this round, each annotator evaluated 50 essays with two feedback comments (a total of 100 essays and 200 feedback comment posts); one GPT-generated and one human-generated. The criteria were the same as those in the training phase, except for one additional question: *"In your opinion, how could this feedback be improved?"*. Figure 1 illustrates an example essay along with its corresponding feedback comments. The essay and the human feedback is retrieved from essayforum[2].

[2]https://essayforum.com/writing/funding-education-healthcare-sectors-may-help-64985/

Criteria	Human Feedback Rank		
	First	Second	Tie
C1 (relevance)	0	1	99
C2 (weakness)	44	13	33
C3 (strength)	8	46	36
C4 (specificity)	45	12	33
C5 (helpfulness)	38	12	40

Table 3: Number of times human feedback was ranked first or second by an annotator, and the number of times there was a tie (out of 90, compared with GPT-3.5 rubric-based) for each criterion.

3.1. Results

Ranking results are available in Table 3. The general trend we observed was that GPT is better at highlighting strengths (C3), while humans are better at highlighting weaknesses (C2). Also, human feedback comments take the lead on being more specific and actionable (C4). The next section will cover our findings in more detail. We would like to make a note that *tie* does not mean the comments were the same. In most cases, we observed that even though comments were different, they were equally important.

3.2. Discussion

Relevance. We did not observe any hallucinations or irrelevant content in GPT-generated feedback comments. Thus, human feedback and GPT-generated feedback were both always ranked as one, except for one case in which human feedback did not contain any feedback on the essay, but instruction on how the author of the essay should include a complete topic along with the essay. GPT would always generate feedback on the essay, even though it might not be able to evaluate if the essay is exactly in response to the topic or not.

Highlighting weaknesses. Overall, humans were much better at highlighting *the most important* weaknesses. We realized that GPT-generated feedback comments follow a specific pattern: It summarizes the essay in 1-2 sentences, lists some of the grammatical issues, and then suggests including more examples and evidence to strengthen the arguments.

Surprisingly, GPT did fairly well at summarizing the viewpoint of the students even when the writing was weak in terms of grammar and sentence structure. In many cases, GPT was better at providing grammatical corrections. Humans would usually mention a few errors (most likely due to time constraints) but if instructed correctly, GPT can list almost all corrections. That being said, there were a few cases where the GPT corrections did not make much sense:

```
“new delhi” should be capitalized as “New Delhi”
and “clinic” should be capitalized as “clinic”.
```

Lastly, GPT tends to suggest including more examples and elaborations for almost all essays which was unnecessary in many cases. Humans, on the other hand, have a better understanding of what are the most important weaknesses of the essay and what should be prioritized in future writings.

Another noteworthy observation was that, in the case of stronger essays (indicating higher proficiency levels), GPT feedback tends to be general and falls short on pointing out weaknesses. In contrast, humans almost always find points for improvement, and are also superior in determining the essay's relevance and assessing the logical coherence of its argumentative elements.

Highlighting strengths. GPT-generated feedback comments were much more encouraging than most human-generated comments. As discussed previously, humans adopt a direct approach, promptly delving into the weaknesses of the essay. But GPT feedback comments always include a few praise sentences which could yield positive outcomes for the student.

Specificity. This aspect heavily depends on the weaknesses of the essay. GPT feedback comments are better at addressing more surface-level issues. For example, they can list more grammatical corrections than humans. On the other hand, humans can give more specific suggestions when it comes to the organization of the essay, relevance to the topic, and argumentative aspects of the essay. Moreover, human feedback tends to focus on a few major or critical issues, delving into them with greater depth, which is often followed by concrete suggestions on how to alleviate the problem. In contrast, GPT-generated feedback tends to offer more generalized and holistic suggestions for improvement.

Helpfulness. Looking at Table 3, in 44% cases, human feedback was more helpful, but for the rest of the cases, it was a tie or GPT feedback was more helpful. This shows that both feedback comments could be helpful despite their respective strengths being centered on distinct perspectives.

Other observations. In a few cases, it became apparent that the content of the essay was biased or included a misconception. Unfortunately, these were not pointed out to the student by either human or GPT. Hence, we think online resources

should be used with caution as they do not always provide the necessary quality control and oversight required for academic purposes.

4. Conclusion

In this paper, we studied and compared two online resources available to students practicing writing skills: community forums, and AI tools such as ChatGPT. We conducted a detailed human evaluation of feedback comments from these two resources, considering aspects that are important in writing constructive feedback. Our study shows that while ChatGPT feedback is more encouraging and positive overall, and in many cases includes appropriate corrections and suggestions, humans are better at giving more specific and actionable comments focusing on the most important issues in the essay.

5. Ethical Statement

We acknowledge the potential for representational harm, a complex issue that is often challenging to quantify. Biases may originate from multiple sources, including annotators, system designers, and the data itself, and these biases can impact how educators interpret students' essays and provide feedback. We are also aware of the documented biases in language models such as GPT. These biases have the potential to inadvertently appear in the outcomes of our study, potentially perpetuating and exacerbating inequalities.

To mitigate such harm, we want to stress that online community forums and AI tools like GPT should be regarded as valuable supplements rather than substitutes for human guidance and expertise in upholding the integrity of scholarly work.

6. Limitations

Future research could consider additional factors in similar studies. Notably, exploring the impact of students' language proficiency levels and writing skills on their perceptions and utilization of ChatGPT feedback would be a valuable endeavor.

It is also important to acknowledge that community forum users sometimes lack the experience and expertise of professional educators, and, in some instances, students provide feedback to their peers on these forums. However, it is essential to underscore that the primary focus of this study was to compare freely available resources for writing practice. It is conceivable that the results may differ if feedback from experienced educators were employed in the assessment.

7. Bibliographical References

Perpetual Baffour, Tor Saxberg, and Scott Crossley. 2023. Analyzing bias in large language model solutions for assisted writing feedback tools: Lessons from the feedback prize competition series. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 242–246, Toronto, Canada. Association for Computational Linguistics.

Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. 2022. AEG: Argumentative essay generation via a dual-decoder model with content planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5134–5148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. Think you have solved direct-answer question answering? Try ARC-DA, the direct-answer AI2 reasoning challenge. *arXiv preprint arXiv:2102.03315*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*.

Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Oeistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, et al. 2023. On the application of large language models for language teaching and assessment technology. *arXiv preprint arXiv:2307.08393*.

Asa Cajander, Mats Daniels, Anne Kathrin Peters, and Roger McDermott. 2015. Critical thinking, peer-writing, and the importance of feedback. *Proceedings - Frontiers in Education Conference, FIE*, 2015-February(February).

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *ArXiv*, abs/2006.14799.

Jack Ende. 1983. Feedback in clinical medical education. *The Journal Of The American Medical Association (JAMA)*, 250(6):777–781.

Juan Escalante, Austin Pack, and Alex Barrett. 2023. Ai-generated feedback on writing: insights into efficacy and enl student preference. *International Journal of Educational Technology in Higher Education*.

Noura Farra, Swapna Somasundaran, and Jill Burstein. 2015. Scoring Persuasive Essays Using Opinions and their Targets. *10th Workshop on Innovative Use of NLP for Building Educational Applications, BEA 2015 at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2015*, pages 64–74.

Kai Guo and Deliang Wang. 2023. To resist it or to embrace it? examining chatgpt's potential to support teacher feedback in efl writing. *Education and Information Technologies*.

Yann Hicke, Abhishek Masand, Wentao Guo, and Tushaar Gangavarapu. 2023. Assessing the efficacy of large language models in generating accurate teacher responses. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 745–755, Toronto, Canada. Association for Computational Linguistics.

Xinyi Huang, Di Zou, Gary Cheng, Xieling Chen, and Haoran Xie. 2023. Trends, research issues and applications of artificial intelligence in language education. *Educational Technology and Society*, 26(1):pp. 112–131.

Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander Fabbri, Yejin Choi, and Noah A. Smith. 2022. Bidimensional leaderboards: Generate and evaluate language hand in hand. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3540–3557, Seattle, United States. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9).

James H. McMillan. 1987. Enhancing college students' critical thinking: A review of studies. *Research in Higher Education*, 26(1):3–29.

Ahmad Omer and Mohhamed Abdularhim. 2017. The criteria of constructive feedback: The feedback that counts. *Journal of Health Specialties*, 5(1):45–45.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Martha N. Ovando. 1994. Constructive feedback: A key to successful teaching and learning. *International Journal of Educational Management*, 8(6):19–22.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. In *Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*.

Nang Kham Thi and Marianne Nikolov. 2022. How teacher and grammarly feedback complement one another in myanmar efl students' writing. *The Asia-Pacific Education Researcher*.

Joshua Wilson and Amanda Czik. 2016. Automated essay evaluation software in english language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers and Education*, 100:94–109.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop*

on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 610–625, Toronto, Canada. Association for Computational Linguistics.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.