Chatbot Feedback on Students' Writing: Typology of Comments and Effectiveness

Besma Allagui(✉)

General Education Department, Rabdan Academy, Abu Dhabi, UAE
ballagui@ra.ac.ae

Abstract. Providing feedback is time-consuming. A chatbot has the potential to facilitate provision of detailed feedback saving instructors time and energy. However, the effectiveness of chatbot feedback has not been determined. In this study, chatbot end-comments on 28 students' written papers were classified by topic and function. The comments were further evaluated for whether they reflected the five characteristics of effective feedback. Results demonstrated the ability of the chatbot to provide a variety of comments. The characteristics of these comments nevertheless, did not seem to reflect feedback best practices. The study provides recommendations for further chatbot feedback research, and in particular calls for more partnership between artificial intelligence developers and writing researchers.

Keywords: First Keyword · Second Keyword · Third Keyword

1 Introduction

From a social perspective, teacher written feedback is the method that teachers-as-readers use to interact with students. It is also an essential component of the writing process that, if well-understood, helps students to improve their writing skills [1]. It draws students' attention to understand their writing weaknesses and regulate their learning [1, 2]. However, providing feedback on student writing is complex and time consuming. It is stated that teachers spend 15–30 min on a single paper [3]. One way to reduce the burden on teachers is through the use of artificial intelligence (AI). Currently, chatbots, a machine learning and artificial intelligence tool which can interact with a user by asking questions and responding to questions, are becoming popular in the writing classroom that may potentially provide teachers with a tool to produce feedback. Researchers have used chatbots to perform various teaching activities including, delivering writing instructions, scaffolding of argumentative writing, providing a conversational partner [4–6]. These studies have demonstrated the promising effects of chatbot on improving students' writing skills. Although chatbots can support writing instruction, it remains to be seen whether chatbots can effectively assist teachers in the provision of corrective feedback.

Given that chatbots are becoming more technically advanced such as ChatGPT, assessing the types and effectiveness of feedback generated by a chatbot could be a necessary to help instructors provide AI-generated feedback more effectively. Feedback can

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14112, pp. 377–384, 2023.
https://doi.org/10.1007/978-3-031-37129-5_31

be classified by its focus (local vs global aspects), by its topic (mechanics, organization, content), or by its role (evaluation, suggestion, reader response) [7]. To date, research has not examined specifically, what type of feedback can be produced by a chatbot and if this feedback reflects the characteristics of effective feedback.

According to writing researchers, good feedback practices should include 5 characteristics: (1) Effective feedback should focus on both content and grammar. Research demonstrated that many teachers currently focus on mechanics which results in improving the readability of the paper without elaborating on its content [7]. (2) Effective feedback is both positive and negative. Feedback should include both praise and criticism. Research revealed that providing students with a combination of negative and positive feedback leads to better writing quality [8]. On the flipside, excessively negative feedback can result in feeling less motivated to rewrite [9, 10]. (3) Effective feedback is specific and relevant to the assignment or task at hand. Researchers found that students are appreciative of detailed and specific feedback because it helped them understand the errors that they must avoid [11] (4) Effective feedback is appropriate. Teachers should focus on what the students can do. It is important that feedback is at an appropriate level of challenge so that student can understand it and make use of it [12] (5) Effective feedback engages learners in the writing process. It is important that teachers do not appropriate students' writing by correcting every single mistake but rather by pointing at inconsistencies and giving them responsibility to make the last decision.

Therefore, in this study we aimed to investigate the types of end-comments and the extent to which these comments reflect best feedback practices. End-comments are described as the paragraph at the end of the paper where the teacher gives summary feedback [13]. We focused particularly on these questions:

1. What types of end-comments does the chatbot provide?
2. Do the chatbot end-comments reflect the feedback best practices?

2 Methodology

The Institutional Review Board determined the study to be exempt from full review. The method employed was exploratory and descriptive. We aimed to identify the types of comments generated by a highly advanced chatbot (ChatGPT) and examine whether the generated comments reflect the feedback best practices as determined by the review of the literature. Two instruments were used to examine the types of comments that AI may produce and the extent to which these comments reflect best teacher feedback practices: students' written samples and evaluation rubric. The written samples were selected from an undergraduate General Education course which focuses on writing, reading and critical thinking skills. The topics of the essay varied between crime and punishment, homeland security and disaster management and military. The students were required to choose their own topic and write a 750 argumentative essay in support of their own position while acknowledging the opposition. Argumentative writing was chosen because it is important for college readiness [14].

First, we asked ChatGPT to read each student essay and write an end-comment that should be specific, detailed, balanced, and engaging. The prompt also asked the chatbot to start with a positive evaluation of the entire paper and organization followed with a

negative statement on style and correctness before ending with a suggestion for the next paper and offering assistance.

After collecting the end-comments, 2 writing teachers read the papers and scored each comment using a rubric specifically designed to for this study based on a review of the literature to measure the extent to which AI-generated comments can reflect best feedback practices. The rubric was reviewed by experts and the final version included 5 criteria pertaining to good feedback: focus, balance, specificity, appropriateness, and learner engagement. Each criterion was measured on a scale from 1 (Never) to 4 (Always) as follows. The raters were trained in one session on using the rubric:

1. Feedback Focus:

   Always: Feedback consistently focuses on both content and mechanics.
   Sometimes: Feedback occasionally focuses on both content and mechanics.
   Occasionally: Feedback focuses on content, but does not provide information on mechanics.
   Never: Feedback does not mention both content and mechanics.

2. Feedback Balance:

   Always: Feedback is both positive and negative.
   Sometimes: Feedback is sometimes both positive and negative.
   Occasionally: Feedback provides an imbalanced view, either only mentioning the strengths or only focusing on the weaknesses.
   Never: Feedback does not mention both the strengths and weaknesses of the situation.

3. Feedback Specificity:

   Always: Feedback provides specific and concrete examples to support the feedback provided.
   Sometimes: Feedback provides some specific examples, but may also rely on generalizations.
   Occasionally: Feedback does not provide specific examples to support the feedback.
   Never: Feedback is solely based on generalizations and does not provide any specific examples.

4. Feedback Appropriateness:

   Always: Feedback is appropriate and clearly stated.
   Sometimes: Feedback may be appropriate in some instances, but may also be vague.
   Occasionally: Feedback is inappropriate and vague.
   Never: Feedback is not appropriate and not clear.

5. Feedback Engagement:

   Always: Feedback encourages engagement and promotes active reflection and improvement.
   Sometimes: Feedback may encourage engagement, but may also be perceived as passive or disinterested.
   Occasionally: Feedback does not encourage engagement and may be perceived as dismissive or disinterested.

Never: Feedback does not encourage engagement and actively discourages reflection and improvement.

Next, comments were read and analyzed qualitatively using a coding scheme adapted from Smith [13]. It was decided to focus on the topic and function of the comment because it was important to see if the AI tool can diagnose writing issues in different aspects and if it is able to adopt different roles.

Data were analyzed both qualitatively and quantitatively. Qualitative analysis included coding the resulting comments on the basis of topic and function using the coding scheme. Quantitative analysis included descriptive and inferential statistics.

3 Results

The produced essays were generally between 603 and 750 words. Results from the thematic analysis of the produced comments are presented in Table 1:

Table 1. Typology of end-comments produced by Chabot

Type	Sub-type	Frequency	Percentage
Topic	Mechanics	30	17.64
	Organization	56	32.94
	Content	84	49.41
Total		**170**	
Function	Evaluation	64	71.9
	Suggestion	25	28.08
	Reader response	0	0
Total		**89**	

The chabot produced the most comments on content especially with regards to introducing the topic and claim and providing counter-arguments. The chabot also provided more of the comments on organization rather than on mechanics. Below is an example of the produced end-comment:

The essay provides a comprehensive overview of the topic of introducing a tax system in the UAE. The writer has discussed the various reasons why the government should start collecting more taxes and the possible challenges that could arise. The essay also includes alternative opinions on the topic, making it well-rounded. The writing style is clear and concise, and the arguments are well-structured. However, some of the sentences could benefit from a more concise phrasing. Overall, the writing quality is good, and the essay is well-suited for its purpose.

In this example, the chabot focused on the content and structure of the essay, including clear and concise writing, well-structured arguments, and the need for some sentences to be phrased more concisely.

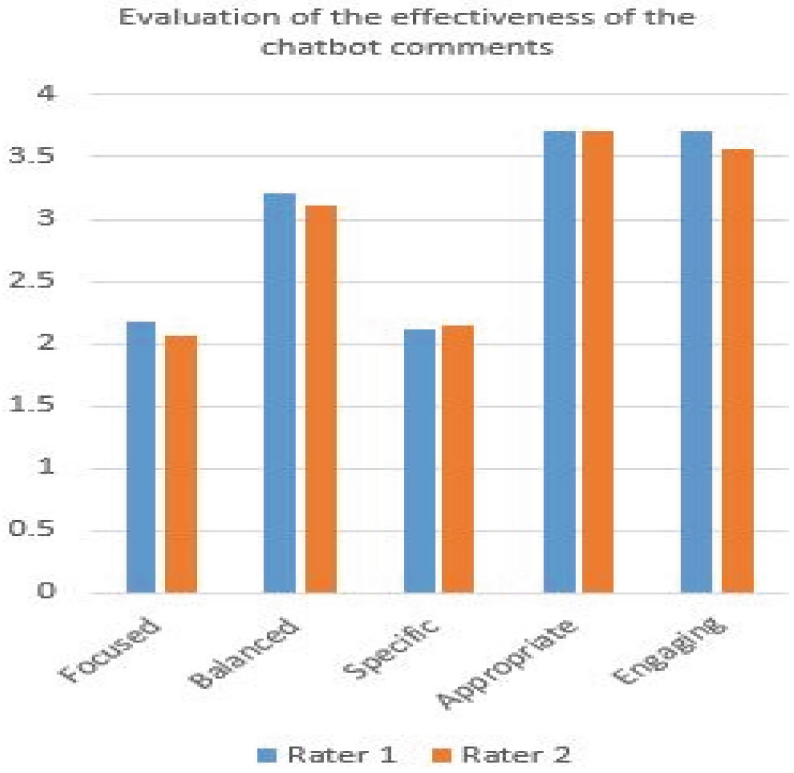

Fig. 1. Human evaluation of the chabot end-comments

The raters' evaluation of the chatbot comments are displayed in Fig. 1:

Figure 1 provides an overview of the ratters' evaluation of the effectiveness of the chatbot comments. Generally, the feedback was found to be appropriate and highly engaging. Raters agreed that the provided comments were balanced including both negative and positive statements. There was also a consensus on the focus and specificity of comments. The comments were general suggestions for improvement rather than detailed feedback (e.g. The language used in the essay can be improved by using more formal and precise language, avoiding repetitions, and using more varied vocabulary).

The interrater agreement for each criterion was measured using Cohen's Kappa. Values ≤ 0 indicate no agreement and 0.01–0.20 indicate none to slight, 0.21–0.40 indicate fair, 0.41– 0.60 indicate moderate, 0.61–0.80 indicate substantial, and 0.81–1.00 indicate almost perfect agreement. The results from the evaluation of the produced comments shows a good inter-rater agreement level between the raters in their assessment of the effectiveness of the feedback on all five criteria ranging from 0.65 for specific to 0.92 for focused.

4 Discussion

The study aimed to examine the types of end-comments that can be produced by a chabot as well as their effectiveness. The results showed that the majority of end-comments are evaluative, including both praise and criticism. While some comments include suggestions for improvement, none of the comments provided reader response. It was also shown that end-comments focused primarily on the content and to a lesser degree on structure followed by mechanics. While this could be attributed to the differences between the samples in terms of content and structure, the results seem to indicate the ability of AI to produce varied comments. The major limitation of the Chabot appeared mainly in not providing reader response such as reactions of whether the writing was interesting, exciting or confusing. Nevertheless, evaluation of the end-comments against the criteria of effectiveness showed that most comments can be judged as highly effective given the high scores provided by both evaluators.

Our findings are in line with the results from [14] study who found that both ChatGPT and Blender are far behind humans in pedagogical abilities. These chatbots need to learn to be emphatic to maximize their usefulness. In our study, we demonstrated that ChatGPT does not provide reader response which serves as a connection between teacher and student and invites students to view their writing from the perspective of readers [3]. Furthermore, results about the extent to which the chatbot comments reflect best teacher feedback practices revealed that using chatbot can provide generic information to help students improve the content, organization and style of their writing which could result in better writing quality papers. The chatbot need to provide more explicit feedback taking into account the students' experience. However, the agreement between the raters that the chatbot-produced comments are balanced and quite engaging seem promising. Evaluators' positive assessment of chatbot comments may mean that AI can greatly contribute to writing instruction and assessment by providing honest praise. Furthermore, acceptance of AI-produced feedback will eventually lead instructors to adopt the new methodologies in their classrooms for better learning and assessment opportunities. These results are supported by [15] who found a high agreement between chatbot grading and human grading. Our study demonstrated that chatbot has more potential than automated tools for providing feedback which has yielded mixed results [16].

5 Implications, Limitations and Conclusion

The study found that chatbot may not be a very effective way to provide detailed and specific end-comments. Therefore, AI should partner with writing researchers to implement the characteristics of good feedback into their language models. In particular, the chatbot needs to improve its social presence and provide more accurate and specific comments. It would be a great idea if the chabot was more intelligent. Meanwhile, chatbot can still be considered a very useful tool in providing students with directions and timely generic suggestions and supporting teachers in assessing students' writing. Overall, the study has contributed to the body of research related to the social aspects of chatbots by examining its potential as a method of providing feedback. Nevertheless, this study

had several limitations. This study is a first step towards examining the effectiveness and typology of chatbot-produced feedback. It suffers from some limitations that should be addressed in future studies. First, the study relied on a small sample of essays on a particular genre. Future studies may include a larger sample written on a variety of genres (narrative, informative, persuasive). Another threat to the generalizability of the current study relates to the prompt. While we asked the chatbot to answer the same question when providing feedback on the students' papers, disclosing more information about each paper may yield different results. Therefore, researchers should continue to explore the effectiveness of chatbot by varying the way the feedback is generated (i.e. reformulating the request). Despite these limitations, the study revealed important features of chatbot-generated end-comments. Improving these features should be a continuous goal for artificial intelligence developers.

References

1. Graham, S., Harris, K.R., Santangelo, T.: based writing practices and the common core: meta-analysis and meta-synthesis. Elem. Sch. J. **115**(4), 498–522 (2015)
2. Black, P., Wiliam, D.: Assessment and classroom learning. Assess. Educ. Princ. Policy Pract. **5**(1), 7–74 (1998)
3. Hyland, K., Hyland, F.: Feedback on second language students' writing. Lang. Teach. **39**(2), 83–101 (2006)
4. Guo, K., Wang, J., Chu, S.K.W.: Using chatbots to scaffold EFL students' argumentative writing. Assess. Writ. **54**, 100666 (2022)
5. Lin, M.P.C., Chang, D.: Enhancing post-secondary writers' writing skills with a chatbot. J. Educ. Technol. Soc. **23**(1), 78–92 (2020)
6. Vázquez-Cano, E., Mengual-Andrés, S., López-Meneses, E.: Chatbot to improve learning punctuation in Spanish and to enhance open and flexible learning environments. Int. J. Educ. Technol. High. Educ. **18**(1), 1–20 (2021). https://doi.org/10.1186/s41239-021-00269-8
7. Chou, C.-Y., Zou, N.-B.: An analysis of internal and external feedback in self-regulated learning activities mediated by self-regulated learning tools and open learner models. Int. J. Educ. Technol. High. Educ. **17**(1), 1–27 (2020). https://doi.org/10.1186/s41239-020-00233-y
8. Wen, Y.: Teacher written feedback on L2 student writings. J. Lang. Teach. Res. **4**(2), 427 (2013)
9. Barkaoui, K.: Teaching writing to second language learners: insights from theory and research. TESL Report. **40**, 14 (2007)
10. Junining, E.: A critique on giving feedback for English as a foreign language (EFL) students' writing. Asian J. Educ. E-Learn. **2**(1) (2014)
11. Zheng, Y., Yu, S.: Student engagement with teacher written corrective feedback in EFL writing: a case study of Chinese lower-proficiency students. Assess. Writ. **37**, 13–24 (2018)
12. Wulandari, Y.: Effective feedback to improve students' writing skills. Educalitra Engl. Educ. Linguist. Lit. J. **1**(1), 10–17 (2022)
13. Smith, S.: The genre of the end comment: conventions in teacher responses to student writing. Coll. Compos. Commun. **48**(2), 249–268 (1997)
14. Tack, A., Piech, C.: The AI teacher test: measuring the pedagogical ability of blender and GPT-3 in educational dialogues. arXiv preprint arXiv:2205.07540 (2022)
15. Ndukwe, I.G., Daniel, B.K., Amadi, C.E.: A machine learning grading system using chatbots. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS (LNAI), Part II, vol. 11626, pp. 365–368. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-23207-8_67

16. Qu, W.: Research on the application of automatic scoring system in college English writing. In: 2016 International Conference on Economics, Social Science, Arts, Education and Management Engineering, pp. 694–698. Atlantis Press, August 2016