International Educational Research; Vol. 7, No. 3; 2024
ISSN 2576-3059 E-ISSN 2576-3067
https://doi.org/10.30560/ier.v7n3p1

Large Language Model for Assignment Feedback on Open-ended Subjective Questions

Zhiqing Liu[1] & Ming Li[1]

[1] Key Laboratory of Educational Informatization for Nationalities (Yunnan Normal University), Ministry of Education,China

Correspondence: Zhiqing Liu, Yunnan Normal University, Kunming, China. E-mail: 1351442105atqq.com

Received: November 13, 2024; Accepted: December 3, 2024; Online Published: December 4, 2024

The research is financed by Graduate Research Innovation Fund Project of Yunnan Normal University in 2024("Research on Exercise Recommendation System Based on Knowledge Tracing", M-B131; "Construction and Application Research of Evaluation Index System for the Application of Big Language Model in Junior High School Information Technology Smart Classroom", M-B135)

Abstract

Feedback is a key factor in motivating and consolidating learning, but in classroom teaching, a teacher needs to provide timely and effective feedback on the homework of dozens of students, which puts much pressure on the teacher. Meanwhile, existing automatic feedback systems are not suitable for open writing tasks. The emergence of ChatGPT has attracted the attention of researchers. We selected 28 open-ended subjective question assignments to input into ChatGPT, and compared and analyzed the scores and comments generated by ChatGPT with those of teachers, to explore the feasibility of using a Large Language Model to provide timely feedback for open-ended subjective questions. Our research indicates that ChatGPT can score and evaluate learners' open-ended subjective homework, and the rating of ChatGPT can be similar to that of teachers. Moreover, ChatGPT's comments can be similar in terms of focus and emotional bias to those of teachers, indicating that ChatGPT's ratings and comments have high credibility.

Keywords: Feedback Generation, Automated Feedback, Large Language Model

1. Introduction

1.1 Introduce the Problem

Homework feedback is an indispensable part of the teaching process, which can help students understand their learning, find their shortcomings, correct their mistakes in time, improve their learning efficiency and performance, and it is also one of the important ways for teachers to understand students' learning situation, which can provide a reference basis for teaching and promote the improvement of teaching quality. However, in the case of classroom teaching, a teacher needs to give timely and effective feedback on the assignments of dozens of students, which is a great pressure on the teacher.

Some researchers have proposed the use of automated feedback systems (AFS) to provide timely assessment of students' work. For example, Marwan et al. (2020) used an automated feedback system for programming topics to provide high school students with immediate feedback on their assignments and found that student's use of this programming environment with immediate feedback was effective in improving their engagement and performance. Some researchers have also used automated writing evaluation (AWE) systems to provide feedback on students' English essays, and these systems focus on grammatical errors and spelling mistakes in students' English essays (Escalante et al., 2023). These automated feedback systems do have good performance in dealing with objective questions and subjective questions with fixed and single answers and rules, but they are not capable of dealing with flexible and open-ended subjective questions, such as case study, opinion statements, research designs, etc. The reason for this is that these automated feedback systems are unable to comprehend the content and ideas of the assignments well. Open-ended subjective questions, on the other hand, have obvious advantages in improving students' critical thinking and innovation abilities, and their proportion in assignments is gradually increasing, requiring teachers to spend more time and energy to provide students with timely feedback on assignments.

Obviously, the problem of the automatic feedback system's understanding of the content and ideas in the homework text must be solved for the automatic feedback system to be able to realize feedback on open-ended subjective questions. Recently, various Large Language Models represented by ChatGPT have gained explosive development, and Large Language Models can learn rich semantic information through a large amount of training data, to more accurately capture the deeper meanings and potential associations of the text. This suggests that Large Language Models have the potential to provide timely feedback on open-ended subjective questions.

In this paper, we aim to explore the feasibility of using Large Language Models to provide timely feedback for open-ended subjective questions. Various studies have affirmed the readability of the text output from the Large Language Model, but readability is only the first condition of feedback, and it is more critical that the feedback provided by the Large Language Model is credible. A simple and direct test is to compare the difference between teacher feedback and feedback from the Large Language Model. Therefore, this paper will test this difference through the following three experiments:

(1) Comparing the difference between multiple feedbacks on the same assignment from the Large Language Model.

(2) Comparing the difference in grading between teacher feedback and feedback from the Large Language Model.

(3) Compare the correlation between teacher feedback and Large Language Model feedback on rubrics.

The experimental object chosen for this paper is the final assignment of 28 students in a class at Yunnan Normal University, which requires students to write an analysis report on the analysis and evaluation of the teacher's language use in the assigned teaching case (Guess How Much I Love You). This assignment required students to have the ability to analyze the case and articulate their viewpoints and was a standard open-ended subjective assignment.

For the first experiment, we will use the Large Language Model to grade and comment on the 28 assignments on three different devices and divide the graded and commented data into three groups, after which we will calculate whether there is a significant difference between the graded data of the three groups, and use the semantic similarity algorithm in the HanLP toolkit to calculate whether there is a significant difference between the three groups of commented data and whether there is a significant difference between the three groups of commented data. " between the three sets of rubric data.

For the second experiment, we will calculate whether there is a significant difference between the teacher's ratings and the 3 sets of rating data separately.

For the third experiment, we will use the semantic similarity algorithm in the HanLP toolkit to calculate whether there are significant differences between the teacher ratings and the three sets of ratings.

By analyzing the experimental data, we conclude that for the same assignment, ChatGPT scores can be similar to the teachers' scores, and ChatGPT comments can be similar to the teachers' comments in terms of focus and emotional bias, which indicates that the ChatGPT scores and comments have a high degree of credibility.

1.2 Automatic Feedback System

Many studies (Higgins et al., 2002; Evans, 2013; Wu& Chang, 2023) have proved that timely and effective feedback is helpful for learners to improve their learning, and this feedback can be categorized into automatic feedback and manual feedback. Automatic feedback is generated through an automatic feedback system and manual feedback is generated through teachers. Under the conditions of classroom instruction, there is some difficulty in providing timely feedback to learners through teachers, so automatic feedback systems have received more and more attention.

Analyzed from the point of view of usage, the existing automatic feedback systems are mainly divided into two categories: one is used for programming practice feedback and the other is used for foreign language writing feedback. Programming practice feedback is used to provide feedback on learners' programming assignments, and these feedbacks include judgments on the correctness or incorrectness of the answers submitted by learners, reasons for the errors, and learning suggestions (Keuning et al., 2018). Foreign language writing feedback is used to provide feedback on learners' foreign language compositions, which can be categorized into four levels according to the complexity of the feedback results: (1) providing only the correct answers; (2) finding errors after comparing learners' inputs word by word with the correct answers stored in the system; (3) predetermining possible errors, storing information related to them in the composition, and presenting them once they match; (4) analyzing the target language's grammatical and lexical rules of the target language and use them as the basis for linguistic analysis of the students' compositions, and mark the problematic parts.

From a technical point of view, existing automatic feedback systems usually generate feedback through two methods: one is to use pre-designed rules, and the other is to use natural language processing techniques. These pre-designed rules are usually designed by experts in various fields, taking into account factors such as the learner's learning behavior (Pardo et al, 2018; Narciss et al,2014), and the difference between the learner's answer and the correct answer (Nagata et al, 2014; Kim et al, 2016; Price et al, 2017).OnTask (Pardo et al, 2018) is an automated feedback system that uses rules to generate feedback, in which the instructor evaluates the learner's learning behaviors, such as attendance, number of homework submissions, homework submission times, assignment grades, etc. Rules are designed to provide feedback to the learners. Apex (Kim et al, 2016) is an automated feedback system for programming exercise feedback, which compares the learner's answers with the standard answers and compares the errors of other learners to give feedback that contains the rating and the reason for the error. Obviously, these systems based on pre-designed rules for feedback have an inescapable drawback; they have to rely on explicit rules or a single standard answer to work, which results in them often appearing to be incompetent in the face of open-ended text assignments. With the development of natural language processing technology, more and more researchers are focusing on using this technology to analyze open-ended text assignments and give feedback. Researchers have fine-tuned various pre-trained language models, such as BART (Lewis et al, 2019) and the GPT family (Radford et al, 2018), to enable the models to perform the complex task of giving text feedback. For example, Jia et al. (2022) designed a BART-based Insta-Reviewer that automatically generates instant feedback on student reports. Li et al. (2021) used GPT-2 to give course feedback to learners, which improved learner engagement.

1.3 Semantic Similarity Calculation

Calculating semantic similarity between texts is research that has attracted much attention in the field of natural language processing, and information retrieval, text categorization, sentiment analysis, machine translation, question and answer systems, etc. are all based on semantic similarity calculation. The earliest semantic similarity calculation method indicates the semantic similarity of two sentences by calculating the frequency of using the same words between sentences, such as Word Frequency and Inverse Document Frequency (TF-IDF), but this method only calculates the frequency of the same words appearing, and does not involve the semantic level, such as "I can't drink pure milk" and For example, "I can't drink pure milk" and "I am lactose intolerant" are not the same words, but the semantics are the same. Therefore, the accuracy of this method is low, and it cannot effectively handle complex tasks.

After that, researchers have proposed several different methods: corpus statistics-based, linguistics-based, and neural network-based methods. The corpus-based statistics approach uses the set of words occurring in a sentence as a feature set, and the cosine of the angle between the corpus-based vectors is used to represent the similarity value. LSA (Latent Semantic Analysis) (1997) is a widely used technique in corpus-based statistics, which uses vectors to represent the words and documents and the angle between the vectors The technique uses vectors to represent words and documents, and judges the relationship between words and documents by the angle between the vectors, and maps the words and documents to the potential semantic space, thus removing some "noise" in the original vector space and improving the accuracy.

Linguistic-based methods utilize the semantic relations between words and grammatical components to determine the similarity of sentences.

Neural network-based methods usually consist of two steps: first, a sentence encoder converts each sentence into a vector; then, a classifier receives the vectors of sentence pairs and classifies them. Sentence encoders map multiple sequences of word vectors onto a single sentence vector, and mainly include methods such as Sequential Recurrent Neural Networks (Seq-RNNs), Tree Structured Recurrent Neural Networks (TreeRNNs), and Convolutional Neural Networks (CNNs). Recurrent neural networks can deal with sequences of arbitrary length, and Long Short-Term Memory Networks (LSTMs) can solve the problem of long-term dependency, but LSTMs can only explore the information of linear structures.

The accuracy of semantic similarity algorithms is often affected by the language of the target data. Because the text data used in this study is in Chinese, the HanLP toolkit, which specializes in the Chinese language, was chosen to compute the semantic similarity. HanLP is a natural language processing library developed by a Chinese developer, He Han (hankcs), in 2014. Since its release, HanLP has been continuously updated and optimized with many new features and performance, and the number of stars on Github has exceeded 31,000. HanLP is very popular among mainstream natural language toolkits and has been tested in academia and industry. classification/clustering, information extraction, semantic analysis, and so on. In this study, we will use the semantic similarity calculation function in HanLP.

2. Method

2.1 Dataset

The data selected for this study comes from the final assignments of 28 students from a class at Yunnan Normal University. The assignment required students to write an analysis report focusing on the analysis and evaluation of the teacher's language use in a specified teaching case ("Guess How Much I Love You"). The teacher will grade the students' work based on four criteria: meeting the word count requirement, integrating theory with case analysis, the clarity of the analysis logic, and the fluency of the evaluative statements. The scores will be categorized into five levels based on the quality of the students' responses: 90 points and above, 80 to 89 points, 70 to 79 points, 60 to 69 points, and below 60 points.

This study's data collection is targeted, selecting the final assignments of 28 students from a class at Yunnan Normal University as the research subjects. These assignments revolve around the specified teaching case of "Guess How Much I Love You," with students required to write an analysis report based on the course requirements. The core content of the report is the analysis and evaluation of the teacher's language use in the case. To ensure the objectivity and comprehensiveness of the evaluation, the teacher will score the students' work in detail from the following four aspects:

Firstly, the word count requirement is examined. Word count is a basic indicator for evaluating the completion of students' assignments and ensures that students can fully express their views and analyses. Secondly, the teacher will look at whether the students are able to combine the learned theory with practical case analysis. This is significant for testing the students' ability to relate theory to practice. Thirdly, the logic of the analysis is evaluated. Clear logic is key to a high-quality analysis report and helps to demonstrate the student's organizational and argumentative skills. Lastly, the fluency of the evaluative statements is reviewed. Language expression is an important aspect of evaluating a student's writing ability, and coherent sentences make the report more persuasive.

Based on these criteria, the teacher will categorize the scores into five levels: 90 points and above is excellent, indicating that the students have outstanding performance in all aspects with in-depth analysis and accurate evaluation; 80 to 89 points is good, meaning that the students have a high level of analytical and expressive ability; 70 to 79 points is average, suggesting that the students have basically met the assignment requirements but still have room for improvement in certain areas; 60 to 69 points is passing, indicating that the students have barely met the basic requirements of the assignment but have deficiencies in multiple aspects; and below 60 points is failing, which means that the students have serious issues in completing the assignment and need to strengthen their learning and practice. By grading and categorizing the assignments of these 28 students, this study aims to provide beneficial feedback for both teaching and learning, promoting mutual improvement in teaching and learning.

2.2 Feedback Generation by ChatGPT

ChatGPT can play a role based on the information in the prompts, which will enable ChatGPT to better understand the requirements of the prompts and fulfill the tasks given to it by the user. In this study, we hope that ChatGPT can play the role of a teacher and grade students' work and give comments according to the grading requirements of the assignments. Therefore, we designed the following prompt: "You are a university professor, you will score and evaluate the students' assignments, and the scoring rules are as follows:

(1) 90 and above: combined with the case, able to skillfully use the knowledge gained to find and reasonably analyze the knowledge points in the case, fully integrated with your views, linked to the theories in the course and the problems in the video for Multi-dimensional analysis, clear logic, sufficient arguments.

(2) 80, and 89 points: combined with the case, can be more skillful in using the knowledge learned to find and reasonably analyze the knowledge points in the case, combined with their views, linked to the theory of the course and the problems in the video to analyze logical clarity, the arguments are more sufficient.

(3) 70, and 79 points: combined with the case, can be used to find and reasonably analyze the knowledge points in the case, with their views, linked to the theory of the course and the problems in the video to analyze, logical clarity, and sufficient arguments. Find and reasonably analyze the knowledge points in the case, have their views, and be able to link the theory in the course and the problems in the video to analyze, the logic is relatively clear, and the arguments are relatively sufficient.

(4) 60 and 69 points: combine the case, be able to link the knowledge in the course and the case, and analyze, the logic is relatively reasonable, there are arguments, and there are no structural defects.

(5) 60 points or less: no combination of the case, not being able to link the course's knowledge from the course and the case and analyze it, the logic is not reasonable, there is no argument, and the structure is missing.

Here are the students' assignments for scoring and evaluation: <Enter the text of the students' assignments>". We insert the text of each assignment into the prompt and ChatGPT will output the scoring and rubric. The scoring by the teacher and ChatGPT is shown in Table 1.

Table 1. Teacher and ChatGPT scores

ChatGPT's scores 1	ChatGPT's scores 2	ChatGPT's scores 3	Teacher's scores
85	85	85	87
85	80	85	85
85	85	85	87
90	90	90	85
90	90	90	90
90	90	90	88
90	85	90	87
85	85	85	87
90	90	90	90
85	90	90	87
90	90	90	90
90	90	90	90
90	90	90	90
90	90	90	90
90	90	90	90
85	85	85	87
85	85	90	85
90	95	90	90
85	85	85	87
90	90	85	87
90	90	90	90
85	85	85	85
85	85	85	85
85	85	85	85
85	90	90	87
90	90	90	90
85	85	85	87
85	85	85	85

2.3 Evaluation Methods

In the first experiment, our goal is to analyze the consistency among the three sets of ChatGPT scores. To do this, we will meticulously count the number of assignments where the score brackets assigned by the three different sets of ChatGPT differ. This statistic will help us understand the variations in scoring criteria between different ChatGPT models. Additionally, to explore the correlation between these three sets of scores, we will employ the Pearson correlation test for correlation analysis. When dealing with the calculation of semantic similarity, one challenge we face is that comments are typically composed of multiple sentences. If we input the entire comment sections directly into the HanLP natural language processing tool, the algorithm might fail to accurately capture the main meaning of the text, incorrectly judging the semantic relevance between the two texts as 0. To address this issue, we have adopted a more refined approach: we first break down the comment text into multiple short sentences, then select two sentences with similar content and input them into HanLP separately to calculate their semantic similarity. Given that the length of each comment may vary, we have decided to use the comment with the fewer number of sentences as the benchmark, calculate the semantic similarity for each pair of sentences, sum these similarities, and finally divide this sum by the number of sentences in the shorter comment to obtain the semantic similarity coefficient between the two comments.

In the second experiment, our focus shifts to the comparison between teachers' scores and ChatGPT scores. We will count the number of assignments where there is a discrepancy in the score brackets between the teachers'

scores and the three sets of ChatGPT scores to assess the consistency between ChatGPT's scoring and teachers' scoring. Moreover, to further analyze the correlation between teachers' scores and ChatGPT scores, we will once again employ the Pearson correlation test to examine the degree of association between the two.

The third experiment focuses on the comparison of the content of the comments. Here, we will use the calculation method introduced in the first experiment to compute the semantic similarity coefficient between teachers' comments and the three sets of ChatGPT comments. Through this calculation, we aim to evaluate whether ChatGPT can capture the teachers' intent and expression in generating comments, thereby providing empirical evidence for the application of ChatGPT in educational settings.

3. Results

3.1 The impact of time and equipment

Upon an in-depth analysis of the scoring sheets from teachers and ChatGPT, it can be observed that there are instances where discrepancies in the scoring brackets exist among the three sets of ChatGPT's ratings. Specifically, there are 3 assignments that show inconsistency in the scoring brackets between ChatGPT's Rating 1 and Rating 2; similarly, there are 3 assignments with scoring bracket inconsistencies between ChatGPT's Rating 2 and Rating 3; and there are 4 assignments with scoring bracket discrepancies between ChatGPT's Rating 1 and Rating 3. This phenomenon indicates that there may be certain variations in ChatGPT's assessment of the same assignment under different scoring criteria.

To further explore the relationship between the three sets of ChatGPT's ratings, we introduced Pearson coefficients for analysis. As shown in Table 2, the Pearson coefficients between the three sets of ChatGPT's ratings are relatively high, suggesting a strong linear correlation between them. This indicates that, for the majority of cases, ChatGPT's evaluation of the same assignment is consistent.

Moreover, we analyzed the semantic similarity between the three sets of ChatGPT's ratings, with the results presented in Table 3. It can be seen from the table that the semantic similarity coefficients between each set of ChatGPT's ratings are also high, indicating a high degree of consistency in the semantics of the comments provided under different scoring criteria.

Synthesizing the data from Table 2 and Table 3, we can conclude that despite some variations in the scoring of certain assignments under different criteria, there is a strong correlation between the ratings and comments given by each set of ChatGPT's evaluations. This result suggests that ChatGPT's scoring and commentary on the same assignment are highly stable and reliable, unaffected by factors such as time and equipment. This provides strong support for the application of ChatGPT in the field of education and offers a more objective and fair evaluation reference for teachers and students.

Table 2. Correlation coefficients between ChatGPT scores

	ChatGPT’s scores 1	ChatGPT’s scores 2	ChatGPT’s scores 3
ChatGPT’s scores 1	1	0.74	0.72
ChatGPT’s scores 2	0.74	1	0.74
ChatGPT’s scores 3	0.72	0.74	1

Table 3. Semantic similarity coefficients between ChatGPT rubrics

ChatGPT’s comment 1 with comment 2	ChatGPT’s comment 1 with comment 3	ChatGPT’s comment 2 with comment 3
0.71	0.73	0.75
0.72	0.7	0.68
0.8	0.78	0.7
0.79	0.8	0.78
0.71	0.79	0.75
0.82	0.76	0.75
0.69	0.7	0.71
0.73	0.8	0.76
0.82	0.8	0.82
0.78	0.69	0.74
0.71	0.76	0.71

0.71	0.8	0.77
0.71	0.75	0.74
0.72	0.74	0.74
0.72	0.69	0.77
0.71	0.8	0.73
0.72	0.7	0.71
0.78	0.74	0.76
0.73	0.71	0.78
0.76	0.68	0.75
0.7	0.71	0.76
0.81	0.76	0.74
0.74	0.74	0.76
0.71	0.71	0.75
0.68	0.69	0.73
0.78	0.76	0.73
0.7	0.7	0.73
0.7	0.71	0.74

3.2 Reliability of Ratings

Upon a detailed comparative analysis of the scoring sheets from teachers and ChatGPT, we have noted some discrepancies in terms of scoring consistency. Specifically, when comparing teachers' scores with ChatGPT's Score 1, there were 4 assignments where the scoring brackets did not align; between teachers' scores and ChatGPT's Score 2, the number of assignments with inconsistent scoring brackets increased to 5; and most notably, between teachers' scores and ChatGPT's Score 3, the number of assignments with mismatched scoring brackets reached 6. These figures indicate that while there is a certain number of scoring discrepancies, the overall number of these differences remains within a manageable range.

To quantify the relationship between teachers' scores and ChatGPT's scores, we calculated the Pearson coefficients and presented the results in Table 4. Analyzing the data in Table 4 reveals a significant positive correlation between teachers' scores and the three sets of ChatGPT's scores. This finding is significant as it demonstrates that ChatGPT can effectively mimic teachers' scoring criteria during the evaluation process, providing a reliable auxiliary tool for assignment assessment.

Furthermore, the data in Table 4 suggests that despite the minor inconsistencies in scoring brackets, the overall trend of ChatGPT's scores aligns with that of the teachers. This implies that ChatGPT is capable of making judgments similar to teachers' evaluations for the same assignment, thereby validating the high credibility of the ChatGPT scoring system. This result not only highlights the advantage of ChatGPT in scoring consistency but also offers strong support for educators, indicating that ChatGPT can serve as an effective tool for teaching assessment, helping teachers evaluate student assignments more efficiently and fairly.

In summary, despite some scoring discrepancies, the high correlation between ChatGPT's scores and teachers' scores provides a solid theoretical foundation for the application of ChatGPT in educational teaching. It also suggests that with continuous technological advancements and optimization, ChatGPT is expected to further assist teachers in educational assessment in the future, enhancing the accuracy and efficiency of educational evaluations.

Table 4. Correlation coefficients between teachers' ratings and ChatGPT ratings

	ChatGPT’s scores 1	ChatGPT’s scores 2	ChatGPT’s scores 3
Teacher’s scores	0.74	0.72	0.63

3.3 The Credibility of Comments

As shown in Table 5, we have conducted a detailed comparison of the semantic similarity coefficients between teachers' comments and three sets of ChatGPT-generated comments. By analyzing the data in Table 5, it is evident that ChatGPT's comments are highly similar to teachers' comments in terms of focus and emotional bias, indicating that the comments provided by ChatGPT have a high degree of credibility.

Specifically, the similarities between ChatGPT's comments and teachers' comments are evident in the following aspects: first, the identification of strengths and weaknesses in students' work; second, the guidance of students' emotional attitudes; and third, the direction for improvement in the assignments. These similarities suggest that ChatGPT can effectively understand the intentions behind teachers' evaluations and reflect them in the generated comments.

Furthermore, the data in Table 5 also shows that although there is a certain gap in semantic similarity between ChatGPT's comments and teachers' comments, this gap is not significant. This further demonstrates the high credibility of ChatGPT in generating comments, which can provide valuable assistance to teachers and alleviate the burden of marking homework. However, it should be noted that there may be some discrepancies in the details between ChatGPT's comments and teachers' comments, so in practical applications, it is necessary to make appropriate adjustments based on the specific circumstances of the teachers to fully utilize the potential of ChatGPT in educational teaching.

Table 5. Semantic similarity coefficients between teachers' and ChatGPT rubrics

Teacher's comment with ChatGPT's comment 1	Teacher's comment with ChatGPT's comment 2	Teacher's comment with ChatGPT's comment 3
0.77	0.75	0.65
0.79	0.68	0.7
0.73	0.73	0.73
0.79	0.76	0.72
0.77	0.73	0.7
0.76	0.71	0.79
0.79	0.72	0.69
0.76	0.73	0.68
0.71	0.72	0.71
0.73	0.75	0.66
0.74	0.72	0.72
0.77	0.7	0.73
0.72	0.69	0.66
0.75	0.75	0.67
0.77	0.68	0.69
0.71	0.76	0.78
0.73	0.76	0.66
0.74	0.68	0.78
0.71	0.73	0.72
0.79	0.69	0.77
0.78	0.74	0.67
0.73	0.73	0.75
0.7	0.73	0.73
0.77	0.73	0.67
0.75	0.76	0.69
0.75	0.77	0.8
0.7	0.75	0.7
0.74	0.67	0.72

4. Discussion

Feedback is a crucial factor in motivating and solidifying learning, but in a classroom teaching situation, a teacher is required to provide timely and effective feedback on the work of dozens of students, which can be a significant source of stress for educators. Our research indicates that ChatGPT is capable of scoring and evaluating open-ended subjective assignments from learners, and the scoring by ChatGPT closely aligns with that of teachers. Moreover, the comments made by ChatGPT are similar to those of teachers in terms of focus points and emotional bias, suggesting that the scoring and commentary of ChatGPT are highly credible.

Feedback plays a vital role in the journey of learning, as it not only motivates students to continue improving but also reinforces the knowledge they have acquired. However, in the traditional classroom teaching model, teachers

face a formidable challenge: how to provide timely, effective, and targeted feedback on the work of dozens of students. This requires not only a high level of professional competence from teachers but also a significant investment of time and effort, which greatly increases their work pressure.

Addressing this issue, our research team conducted a series of explorations and found that ChatGPT, an artificial intelligence technology, holds immense potential in the field of education. The study shows that ChatGPT can accurately score and provide in-depth evaluations of open-ended subjective assignments from learners. Surprisingly, the scoring results from ChatGPT closely match those of manual scoring by teachers, which to some extent proves the reliability of ChatGPT in scoring.

Furthermore, we conducted an in-depth analysis of the comments made by ChatGPT and found that they closely resemble teachers' comments in terms of focusing on key knowledge points and emotional expression. This means that ChatGPT is not only able to provide objective scores but also offer comments with a human touch, thereby better guiding students to improve their learning methods and enhance their learning outcomes. This finding further strengthens our confidence in the role of ChatGPT as an educational aid.

In summary, as an efficient artificial intelligence assistant, the application of ChatGPT in homework scoring and evaluation helps to alleviate the workload of teachers and improve the quality of education, while also providing more personalized and timely learning feedback to learners. This offers valuable insights and inspiration for the process of educational informatization and intelligence in our country. Based on this, we look forward to ChatGPT playing an even greater role in more educational scenarios in the future, contributing to the development of education in our nation.

Acknowledgments

This work was supported by Graduate Research Innovation Fund Project of Yunnan Normal University in 2024(“Research on Exercise Recommendation System Based on Knowledge Tracing”, M-B131; “Construction and Application Research of Evaluation Index System for the Application of Big Language Model in Junior High School Information Technology Smart Classroom”, M-B135).

References

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: insights into efcacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, (20), 55. https://doi.org/10.1186/s41239-023-00425-2

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research*, *83*(1), 70-120. https://doi.org/10.3102/0034654312474350

Higgins, R., Hartley, P., & Skelton, A. (2002). The conscientious consumer: Reconsidering the role of assessment feedback in student learning. *Studies in Higher Education*, *27*(1), 53-64. https://doi.org/10.1080/03075070120099368

Jia, Q., Young, M., Xiao, Y., Cui, J., Liu, C., Rashid, P., & Gehringer, E. (2022). Insta-Reviewer: A Data-Driven Approach for Generating Instant Feedback on Students' Project Reports. *International Educational Data Mining Society*.

Keuning, H., Jeuring, J., & Heeren, B. (2018). A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education (TOCE)*, *19*(1), 1-43. https://doi.org/10.1145/3231711

Kim, D., Kwon, Y., Liu, P., Kim, I. L., Perry, D. M., Zhang, X., & Rodriguez-Rivera, G. (2016). Apex: automatic programming assignment error explanation. *ACM SIGPLAN Notices*, *51*(10), 311-327. https://doi.org/10.1145/3022671.2984031

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211. https://doi.org/10.1037/0033-295X.104.2.211

Lewis, M. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*. https://doi.org/10.18653/v1/2020.acl-main.703

Li, C., & Xing, W. (2021). Natural language generation using deep learning to support MOOC learners. *International Journal of Artificial Intelligence in Education*, *31*, 186-214. https://doi.org/10.1007/s40593-020-00235-x

Marwan, S., Gao, G., Fisk, S., Price, T. W., & Barnes, T. (2020, August). Adaptive immediate feedback can

improve novice programming engagement and intention to persist in computer science. In *Proceedings of the 2020 ACM Conference on International Computing Education Research* (pp. 194-203). https://doi.org/10.1145/3372782.3406264

Nagata, R., Vilenius, M., & Whittaker, E. (2014, June). Correcting preposition errors in learner English using error case frames and feedback messages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 754-764). https://doi.org/10.3115/v1/P14-1071

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education*, *71*, 56-76. https://doi.org/10.1016/j.compedu.2013.09.011

Pardo, A., Bartimote, K., Shum, S. B., Dawson, S., Gao, J., Gašević, D., ... & Vigentini, L. (2018). Ontask: Delivering data-informed, personalized learning support actions. *Journal of Learning Analytics*, *5*(3), 235-249. https://doi.org/10.18608/jla.2018.53.15

Price, T., Zhi, R., & Barnes, T. (2017). Evaluation of a Data-Driven Feedback Algorithm for Open-Ended Programming. *International Educational Data Mining Society*.

Radford, A. (2018). Improving language understanding by generative pre-training.

Wu, L. J., & Chang, K. E. (2023). Effect of embedding a cognitive diagnosis into the adaptive dynamic assessment of spatial geometry learning. *Interactive Learning Environments*, *31*(2), 890-907. https://doi.org/10.1080/10494820.2020.1815216