IJCESA
International Journal of Community Empowerment & Society Administration
Volume 3 Issue 1 January 2026/ Paper Id: IJCESA-V3I1P101/Page No: 01-11

Original Article

GenAI for Revolutionising Writing Assessment: A Case Study on the Comparative Advantages of ChatGPT 3.5, ChatGPT Edu, and Human Teachers in Assessing Essays

Md. Saiful Alam[1], A H M Ohidujjaman[2], Md. Nurullah Patwary[3], Humayra Akhter[4]

[1]*Assistant Professor, Department of English, World University of Bangladesh, Uttara 17, Dhaka 1230, Bangladesh.*
[2]*Lecturer, Department of English, World University of Bangladesh, Uttara 17, Dhaka 1230, Bangladesh.*
[3]*Assistant Professor, Department of English, World University of Bangladesh, Uttara 17, Dhaka 1230, Bangladesh.*
[4]*Senior Lecturer, Department of English, World University of Bangladesh, Uttara 17, Dhaka 1230, Bangladesh.*

Received Date: 18 November 2025 Revised Date: 20 December 2025 Accepted Date: 02 January 2026

Abstract: *The historical normativity is that education and AI make a parallel of reciprocal effect systems, where innovations in one field help the other progress. One of the recent developments in scientific innovations relating to writing pedagogy is the emergence of large language models such as ChatGPT. This has created novel prospects for transforming educational practices, particularly in the realm of writing assessment. This development has prompted the researchers to examine how ChatGPT Edu performs over its predecessor, ChatGPT 3.5, in the framework of Automated Essay Scoring (AES). This study specifically evaluates the performance of ChatGPT Edu and ChatGPT 3.5 in comparison to human teachers in assessing student writing. It applies a qualitative case study research design to address the research problems. To collect the required amount of data, the study uses a variety of data sources, such as EFL students' handwritten essays, ChatGPT-produced grades, comments, evaluations, as well as thorough evaluations from an experienced EFL writing teacher. The study applies the summative content analysis approach to analyze the qualitative data. The findings reveal that ChatGPT Edu is superior to ChatGPT 3.5 in all three assessment dimensions, and it can contribute to a higher degree of reliability and pedagogical value in second language writing assessment. The findings also demonstrate that ChatGPT Edu has a stronger capacity to score that closely matches teacher-assigned scores. These findings suggest that the cutting-edge AI tools can be of great support to EFL teachers in assessing writing. Furthermore, this study also contributes to the ongoing discussion on how successfully AI-supported instruments can be integrated into EFL education.*

Keywords: *ChatGPT, ChatGPT Edu, Language Assessment, Human Assessment*.

I. INTRODUCTION

Parallel with the immense influence of AI in contemporary daily life affairs, Artificial Intelligence (AI) emerged with a transformative role in education, and in recent years, the educational stakeholders have proactively adopted AI technologies in writing pedagogy and assessment practices (Tahiru, 2021; Okunlaya, Abdullah & Alias, 2022). Particularly in tertiary education, advanced AI-generated instruments have contributed to the enhancement of more efficient, personalized, and innovative educational practices (Zhai et al., 2021). As a result, the role of AI technology has generated great enthusiasm among the learners and educators. Many have started using this technology, considering that it is useful for teaching and learning. However, this enthusiasm for and use of the AI-supported tools have given rise to some critical debates too, particularly regarding their capacity to remodel assessment practices (Felix, 2020; Schiff, 2022; Michel-Villarreal et al., 2023; Crompton & Burke, 2023; Kayyali, 2024).

A noteworthy development in the use of AI-supported tools is their ability to automatically score essays, which is known as automated essay scoring (AES). In the EFL writing context, the machine learning model of AES has shown outstanding potential (Gardner, O'Leary & Yuan, 2021). As various generative tools, such as ChatGPT, have emerged, research interest has also grown in parallel to explore their effectiveness in various educational contexts (Ali et al., 2023; deWinter, Dodou & Stienen, 2023). Recent innovations show that AI generative instruments like ChatGPT are capable not only of generating written content but also of scoring and evaluating it (Gardner, O'Leary & Yuan, 2021; Ali et al., 2023). However, this machine-led capacity has raised questions about their effectiveness in maintaining accuracy, fairness, and proximity to the standards of human evaluation (Mizumoto & Eguchi, 2023; Latif & Zhai, 2024). Considering these concerns, researchers highlight the need for continual assessment of the AI systems to check if these systems meet the expectations and standards of human teachers (Zhai & Nehm, 2023; Herbold et al., 2023).

The recent developments in AI-generative tools have ushered in significant progress in language teaching activities. One such tool is ChatGPT Edu, which is an advanced version of ChatGPT 3.5. ChatGPT Edu is specially designed for

educational contexts and offers some exceptional services for teaching and learning. However, there is a lack of knowledge in the existing research about the performance of ChatGPT Edu, as the existing research has predominantly explored the performance of its earlier version, ChatGPT 3.5, in assessing writing (Latif & Zhai, 2024; Altamimi, 2023). This gap is especially significant when it comes to its performance in evaluating student writing and its comparison with its earlier version, ChatGPT 3.5, and with experienced human teachers teaching and assessing EFL writing.

The current study aimed to fill this research gap. In order to address this gap, the study conducted a qualitative study to explore comparative performances of ChatGPT 3.5, ChatGPT Edu, and experienced EFL teachers in assessing written content. To explore the comparative performances of the evaluation sources, this research adopted a qualitative case study approach. This approach used Automated Essay Scoring (AES) as its central framework. Based on this framework, this study assessed the performance of ChatGPT Edu and then compared its performance with that of ChatGPT 3.5 and experienced EFL teachers. This research endeavor was guided by three criteria for assessing writing: assessment accuracy, scoring fairness, and proximity to human evaluation standards. Through exploring a nuanced understanding of the performance of the AI tools, this study offers notable primary insights into the practicality and effectiveness of ChatGPT Edu in writing assessment.

A. Objectives and Research Questions

The primary objective of this study is to investigate the comparative effectiveness of ChatGPT Edu in assessing handwritten essays in contrast to its earlier version, ChatGPT 3.5, and human English faculty raters. To achieve this aim, the study addresses the following research question:

How does the automatic essay scoring effectiveness of ChatGPT Edu differ from that of ChatGPT 3.5 and human raters?

B. Literature Review: Theoretical Framework

a) ChatGPT: Emergence and Evolution

This study explores the evolution of ChatGPT, focusing on its introduction, versioning, and updates within the educational sector, rather than its scientific origins and development history. From an educational perspective, ChatGPT represents a recently developed artificial intelligence tool capable of performing multidimensional functions to facilitate academic endeavors. ChatGPT was launched in late 2022 and gained popularity around the globe, particularly within the context of language education. Primarily, it functions as an artificial intelligence system (Rospigliosi, 2023), demonstrating capacities that surpass those of human intelligence in a remarkably natural manner. Moreover, it represents a recent advancement in natural language processing (Shafik, W., 2024). The name chosen by OpenAI for this model, ChatGPT, underscores its focus on chat or conversation. This conversational characteristic of this tool led to its recognition as a conversational AI (Dwivedi et al., 2023).

The initial publicly accessible iterations of ChatGPT were the non-premium versions, ChatGPT 3 and ChatGPT 3.5, whereas its premium counterpart, ChatGPT 4, offered supplementary functional enhancements and was available through subscription-based models (Kayyali, 2024; Chen et al., 2024). It appears that various versions of ChatGPT have already been employed in educational contexts (Adeshola & Adepoju, 2023; Firat, 2023; Su & Yang, 2023; Polverini & Gregorcic, 2024). Certain comparative studies have outlined slight differences in performance metrics between the primary and recent versions of ChatGPT (Abramski et al., 2023; Koubaa, 2023; Scheschenja et al., 2024). A clear understanding of the utility of ChatGPT opens a new avenue for research into its potential applications within learning and teaching communities. Amid this ongoing versioning and re-versioning, the recent launch of ChatGPT Edu has occurred. Consequently, this study aims to explore its potential for integrating ChatGPT's educational version into language assessment practices, specifically grading students' work.

b) ChatGPT Edu: An Introduction

As an endeavor to get introduced to the latest version of ChatGPT, this study first studied the introduction page "Introducing ChatGPT Edu" on the OpenAI website (see here: https://openai.com/index/introducing-chatgpt-edu/). This review helped to understand the special features and exclusive functionalities of ChatGPT Edu. In this introduction, OpenAI describes ChatGPT Edu as its flagship model for educational applications and designates universities as its primary users. OpenAI strongly recommends that academicians use this tool responsibly and advocates for setting up ethical boundaries to guide its use in educational settings. The introduction provides a contextual background by highlighting the significant achievements of ChatGPT's standard version in higher education. It also mentions that many well-known institutions, including Oxford University, the University of Pennsylvania, the University of Texas, Arizona State University, and Columbia University, have successfully implemented it.

In technology, ChatGPT Edu is built on the system of the GPT-4 model (Reddit community post, 2024). This system enables ChatGPT Edu to engage in various academic tasks, such as analyzing written texts, analyzing images, utilizing advanced tools for data analysis, coding, interpreting texts, and solving math problems. These built-in service features of ChatGPT Edu offer services for evaluating students' writing skills. This assumed assessment capacity of ChatGPT Edu opens up the possibility of testing it through empirical studies, which is supported by the existing success stories of ChatGPT in reducing assessment time. The present study diverges from this assumption to explore how ChatGPT Edu distinguishes itself with its enhanced catalytic capabilities and performance in assessing the essays of undergraduate EFL students. This study evaluates the writing assessment efficiency of ChatGPT Edu compared to ChatGPT 3.5 version and human assessors.

c) ChatGPT for Grading and Marking: Assumptions and Evidence

The initiation of ChatGPT has encouraged inquiries into its prospect to facilitate pedagogy, educational learning, and support in assessment practices (Božić & Poola, 2023). One of the major focuses is the applications of ChatGPT in examining students' academic assignments and examination essays (Javaid, Haleem, Singh et al., 2023). Considering this grading functionality of ChatGPT, scholars have colloquially coined it as an "auto-grader" (Chang & Ginter, 2024). Recent research endeavors have focused on empirically investigating ChatGPT's grading proficiency, and these studies have yielded results in three major dimensions. First, existing research reveals that ChatGPT can assess students' assignments and examination essays with a degree of resemblance to human teachers (Jukiewicz, 2024). In addition, its grading services include an explanation of scoring and offering better solutions (Javaid, Haleem, Singh et al, 2023). Based on the set criteria, such as content, styles, and organization, ChatGPT can grade long essays (ibid) and provide remarks, like correct, almost correct, and incorrect (Jukiewicz, 2024). Second, in spite of its grading capacity, ChatGPT has some limitations too, and these need to be considered. One of the particular concerns is whether it is suitable for grading a wide range of student work (Jukiewicz, 2024). The third and most notable dimension is to investigate the impact of GPT advancements (e.g., from GPT 3.5 to GPT 4) on its grading efficacy across various metrics (Chang & Ginter, 2024). This study focuses on the third dimension and examines the efficacy of the latest specialized version of ChatGPT, known as ChatGPT Edu, in evaluating essays written by undergraduate EFL students in comparison to its previous version, GPT 3.5, and human instructors.

d) Human Assessment and ChatGPT Assessment: The Existing Divergences

In assessing students' work, a noteworthy deviation arises between the assessments conducted by human instructors and those facilitated by ChatGPT (Jukiewicz, 2024). This divergence is primarily manifested in the variation of assigned scores. It is observed that ChatGPT frequently awards lower grades than human raters. Second, the time-based dimension of assessment also differs considerably. Human raters have to invest considerable hours in the grading process, while ChatGPT does it in just a few seconds (Jukiewicz, 2024). Third, another significant dimension of grading variation is also observed in the degree of flexibility and strictness. ChatGPT exhibits a rigorous adherence to predefined grading criteria, whereas human raters show more flexibility in their assessments (Jackaria, Hajan & Mastul, 2024). Another remarkable dimension of deviation between these two sources of raters was observed by Steiss et al. (2024). They observe that human grading is influenced by various factors like personal experience, training, pay, and time, while ChatGPT's assessments are independent of these variables.

The present study takes place in the context of a university. It focuses on the methods adopted by novice lecturers in assessing and grading assignments and handwritten exam papers of undergraduate EFL students. The evaluation is prompted by the introduction of an advanced AI-supported grading tool, ChatGPT Edu, an upgraded version of ChatGPT 3.5.

e) ChatGPT Edu as an Educational Breakthrough: A Theoretical Underpinning

The present study is based on Page's theory of Automated Essay Scoring (AES) (Page, E. B., 1966.; Page, E. B., 1968; Page, E. B., 2003). This theory highlights the use of computer-based advanced systems in assessing student writing. AES is a viable alternative to traditional manual grading by teachers, as it offers some specialized services to overcome the limitations of manual grading. Manual essay grading often takes a lot of time, involves cost, and may produce inconsistent results due to grader bias, exhaustion, and unwillingness to give detailed feedback (Zhang & Wang, 2023; Bukowski & Tokowicz, 2021). On the other hand, AES is improved, cost-effective, time-efficient, and consistent in maintaining accuracy (ibid.) in spite of having some inherent challenges of essay assessment, such as pattern recognition and language understanding (Ramesh & Sanampudi, 2022). Therefore, advanced evaluation technologies have created a significant amount of enthusiasm in EFL assessment education. Considering the challenges and difficulties traditional essay grading offers, this study introduces a cutting-edge AI technology, ChatGPT Edu. This AI-supported system is rich in natural language processing (NLP) features, offering fast and convenient essay evaluation. This latest AI-supported system is aligned with the wider objectives of improving the efficiency of assessing student writing.

The Automated Essay Scoring (AES) system is not a fixed method but rather an ever–developing one. It keeps changing and improving in response to growing challenges in the practices of grading and evaluating writing content. As an

advanced evaluation system, ChatGPT Edu offers some innovative features that have improved the efficiency of AES. It aligns with the current direction of the field that exploits the benefits of modern technological advances. Some of the traditional AES systems—such as Intelligent Essay Assessor (IEA) (Foltz, Laham & Landauer, 2003), Project Essay Grader (PEG) (Ajay, Tillett & Page, 1973; Shermis et al., 2001), Bayesian Essay Test Scoring System (BESTY) (Rudner & Liang, 2002), and more recent models such as the Attentional Multi-Reading (AMR) model (Dong, Zhang & Yang, 2024)—were designed to evaluate some key features of student writing. These include grammar, vocabulary, structural and thematic coherence, content relevance (often through latent semantic analysis), style, and overall organization, using methods such as pattern recognition and syntactic/semantic analysis. AES systems are proficient in furnishing both specific evaluations based on predetermined criteria and holistic assessments of essays [40]. The current study investigated the efficiency of ChatGPT Edu, a cutting-edge technology in AES.

C. Literature Review: Previous Studies

In recent years, the use of generative AI tools, such as ChatGPT, has assumed a prominent role in language education, as these tools offer some attractive services like instant and interactive feedback for language development. In particular, these have been extensively adopted in the domain of EFL writing assessment. The growing adoption of AI services, such as ChatGPT, has created both enthusiasm and concern among EFL teachers and learners, prompting research attention into the effectiveness and pedagogical implications of these tools. A significant number of such studies have focused on the comparative performance of these tools in relation to traditional human teachers' evaluation practices.

Lampropoulos and Papadakis (2025) investigated the wider application of AI and social robots in education. The findings identified that these systems can act as intelligent, human-like tutors and offer individualized explanations and feedback about a learner's strengths and weaknesses. However, the authors also stressed the need for further empirical research and the development of updated ethical guidelines before widespread adoption of AI technologies for both personal and institutional use.

With regard to computerized writing assessment, a number of studies have compared ChatGPT's effectiveness with that of human raters. In the Turkish context, Kınık and Çetin (2023) conducted a mixed-methods study that examined whether ChatGPT 3.5 could lessen teachers' workload by assessing student essays. The findings indicated that ChatGPT was able to provide rapid feedback. However, its evaluations often deviated from those of human raters, both in terms of scoring and feedback quality. However, the authors also stressed the need for the development of updated ethical guidelines before extensive use of AI technologies and for strict teacher supervision.

A similar study was conducted by Potchong et al. (2024) in the Philippine EFL context. The study compared the essay evaluation performance of ChatGPT 3.5 with that of human raters. They used a standardized rubric to assess student writing. Although ChatGPT 3.5 demonstrated moderate consistency in its evaluation, it showed poor correlation with those of human raters. The findings indicate that ChatGPT 3.5 lacks the depth and exactness necessary for evaluating student essays and suggest that teachers should be extra cautious when using ChatGPT in rating students' written works. The authors also stressed the need for further research to evaluate the accuracy and practicality of using AI applications in evaluating student essays.

In the African EFL education context, Bouziane and Bouziane (2024) carried out another similar study. The study explored the essay evaluation performance of ChatGPT, an AI-supported tool. 100 undergraduate EFL students from a university in Morocco participated in the study. The study employed both ChatGPT and human teachers to analyze the essays written by the students. The analysis then compared the essay evaluation performance of ChatGPT with that of human teachers. The findings indicate that ChatGPT was effective in assessing some basic aspects of writing but was unable to evaluate more complex elements, such as sustaining thematic consistency and developing well-thought-out arguments. The findings suggest that a mixed-methods approach to assessing student essays, which combines the judgment of human evaluators with AI-generated evaluations, could be very effective.

Finally, Steiss et al. (2024) performed a large-scale study in the school education context of the USA. The study compared formative feedback generated by ChatGPT and human raters across five distinct feedback dimensions. Students in 26 different classrooms (Grades 6–12) from two school districts in Southern California, USA, wrote source-based argument essays in history. For this study, we randomly sampled 200 mixed-ability students from the larger study. The findings indicated that human teachers generally outperformed ChatGPT in most dimensions. As graders, they did better, particularly in delivering accurate, clearly articulated, and useful feedback. The results also showed that ChatGPT was more dependable in aligning its feedback with rubric criteria and meaningfully lessened human graders' evaluation time. The study recommended that ChatGPT could be a useful assistant for EFL teachers, which would immensely help them provide rapid feedback in large classrooms and lessen their workload.

D. Research Gap and Justification for the Research

The current research reviewed in this section suggests that most previous studies have largely examined the effectiveness of various AI-supported essay evaluation tools from the ChatGPT family. The major findings include the challenges related to its scoring consistency, understanding themes, and evaluation quality (3–6). However, there is a substantial dearth of research examining the performance of ChatGPT Edu in comparison to other AI tools. This gap highlights the need for further research that involves the comparative analysis of the effectiveness of ChatGPT Edu with other generative AI tools and human graders in evaluating student essays.

II. MATERIALS AND METHODS

A. Research Design

a) Philosophy and Determination of a Case Study

The present study is based on the constructivist, realist, and relativist assumptions (Harré & Krausz, 1996; Devitt, 1997; Amineh & Asl 2015; Asmawi & Alam, 2024). That is, there can be many subjective experience-based realities around a phenomenon (e.g., human essay scoring vs AI essay scoring). There could be one reality, but it can be explained in varying contexts (e.g., English major, tertiary education, global south, co-evolution of education and AI, etc.). The mentioned confluence of philosophical trio signals that there could be varying contextual determinations of co-evolutionary scenarios in which versions of AI and profiling gaps of human teachers make distinct cases. Therefore, a qualitative case study was designed to gain a nuanced understanding of the phenomenon. It is believed that a case study offers a richer and in-depth exploration of a phenomenon in real contexts (Asmawi & Alam, 2024; Flyvbjerg, 2011). The nuanced understanding offered by the designed case study is constitutive of the composite comparisons of scores made by the respective essay-scoring performances of ChatGPT 3.5, ChatGPT Edu, and human teachers in the given context.

b) Description of the Case

The case constitutes a certain university context, subjects (EFL teachers), AI tools, EFL students' hand-written essays in English, a time frame, and tertiary EFL teaching. The study takes place at World University of Bangladesh, one of the renowned private universities in Bangladesh. More specifically, the case is bound by the English Department of the mentioned university. The uniqueness of the case (i.e., English department context) is that the teachers of the department have the highest level of interest in the area of AI in ELT, compared to their counterparts in the Faculty of Arts and Social Sciences. For instance, they published four research articles on AI in ELT in Scopus-indexed journals in 2024 alone. They have sound AI literacy as well. They have diverse interests and attentional focuses. That is, they are enthusiastically engaged in making a bridge between the research evidence and praxis. They are keenly exploring and seeking a continuous understanding of the cyclical relationships among research, evidence, and practice. There is a collectiveness among the faculty members that is directed toward pedagogical innovation, collaboration, knowledge sharing, and research to develop, diversify, and modernize the ELT service delivered by the department and thus make a difference as a department and an academic team. The faculty members include one professor, five assistant professors, four senior lecturers, and nine lecturers, all of whom have already benefited from their personalized uses of ChatGPT 3.5 for making ELT materials and other scaffolds.

They are all modernist, flexible ELT practitioners and adopters. In addition, they have been manually scoring essays as part of their routine as EFL teachers. Given that background, they are also traditionalist practitioners. To rationalize their adoption or rejection of ChatGPT Edu as an AI-based, easy, and advanced scorer, comparing its performance with that of its preceding AI tool (i.e., ChatGPT 3.5) and human scorers is crucial. This comparison helps decide whether there is an equilibrium or an asymmetry between AI and human scoring and thereby provides the rationale for the case study. There is a contextual flexibility endorsed and encouraged by both the university and the department that allows teachers to exercise professional agency. Therefore, they can make micro-level decisions about personalized accommodation and the adoption of AI for the promotion of academic excellence in the department. The students take closed-book exams, and they handwrite their essays in a restricted time twice a semester (mid-term and semester end). The teachers teach both knowledge courses for English and ELT majors and language courses for business, law, and engineering majors. Their teaching load is rather high, and they have to score a lot of essays in a rather short time. Given this backdrop, the present case study is further rationalized.

B. Data

The research question guiding the present study focuses on the effectiveness of ChatGPT Edu in automatic essay scoring compared to its predecessor, ChatGPT 3.5, and human ELT teachers. Considering the data aspects of scores, essays, and scorers, the present study collected qualitative content data. The data include nine handwritten essays collected from the semester final examination scripts of nine EFL students majoring in Law. As this is preliminary research, focus on ChatGPT Edu, large-scale data collection, and triangulation were not intended. The sampling was based on three quality dimensions of

the essays. Three essays were written by highly proficient (HP) students, three by moderately proficient (MP) students, and three by low-proficient (LP) students. These varying quality levels were selected to bring forth the nuances of performance consistency of the scorers vis-à-vis varying essay qualities. The data profiles are represented in the following table:

Table 1 : Students, Student Types, and Essay Lengths

Student	Student type	Essay length	Score
A	HP	210 words	highest
B	HP	190 words	highest
C	HP	201 words	highest
D	MP	180 words	Moderate
E	MP	167 words	Moderate
F	MP	179 words	Moderate
G	LP	153 words	Lowest
H	LP	147 words	Lowest
I	LP	154 words	Lowest

C. Data Collection Method

The data collection started by obtaining oral consent from the three EFL teachers teaching three sections of law students in the spring of 2024, according to the guidelines of the research and ethics committee of the university the authors belonged to at the time of the study. After they had agreed to collaborate in the data collection process, the researchers asked each of them to provide the researchers with three specimen copies of scripts (one from a highly proficient student, one from a moderately proficient student, and one from a low-proficient student). Thus, nine scripts were collected in the same way, with three from each of the three teachers. The teachers were informed beforehand that their scripts would be re-scored by AI tools. The teachers were given independence to follow the ideal rubrics that they normally use as the internal self-styled benchmark in assessing academic coursework, rather than dictating any externally formulated rubric. Similarly, ChatGPT was given freedom to use its normative training data-based rubrics. This chance-creativity and normativity were aimed at finding out the corresponding salience in the human-ChatGPT map of essay scoring. That is, all the collected scripts were already scored by the teachers in question. The teachers randomly selected three representative sample scripts with the highest, moderate, and lowest scores. The scored scripts were subsequently handed over to the researchers by the EFL colleagues. The scoring rubrics that the scorers used were also requested, and the scoring teachers handed them over to the researchers. Thus, the human scores were collected. Afterward, the essays were imaged and converted into texts by using a Google app. The handwritten essays were turned into a typed format. Subsequently, the typed essays were submitted sequentially to ChatGPT 3.5 and ChatGPT Edu for scoring. For homogeneity and fairness, the prompts were all the same for both versions of ChatGPT. Then, the scores from ChatGPT 3.5 and ChatGPT Edu were extracted from the conversation prompts, recorded in a separate data collection folder, and saved for further use. Upon completion of scoring of all nine essays, the data collection stage closed.

D. Data Analysis Method

The present study is, by nature, qualitative and framed into a case study paradigm. Therefore, the findings are not essentially suited to bring forth a quantitatively generalized truth. The small-scale data were typologically qualitative. However, the qualities of the essays were put into numerical values for analytical purposes, rather than for the application of descriptive or inferential statistical procedures. A manual comparative analysis framework was adopted to compare the three scores generated by ChatGPT 3.5, ChatGPT Edu, and human teachers (Slinkard & Singleton, 1977). Under this framework, the data analysis was phased into three stages: tabulating scores, identifying patterns, and categorizing differences. The scores were organized in a table under various category variables, including essays, ChatGPT 3.5 scores, ChatGPT Edu scores, human scores, and score differences (ChatGPT-human and ChatGPT Edu-human). The next stage of pattern identification followed the presentation of tabular data. At this stage, the consistency of the compared scores was identified. This included those whose scores are consistently higher than those previously identified. Any outliers in the scores (whether unusually high or low) were identified and analyzed. In the final stage of categorizing the differences, the final findings came up based on the magnitude and direction of the differences (e.g., minimal, moderate, and large deviations). Additionally, the comparisons between the nuances of the rubrics of ChatGPT 3.5, ChatGPT Edu, and human scorers were also accomplished by categorizing and identifying the composites of those rubrics.

III. RESULTS AND DISCUSSION

A. Results

The present study addressed the research question of "How does the automatic essay scoring effectiveness of ChatGPT Edu differ from that of ChatGPT 3.5 and human raters?" The findings offer the following primary insights:

As indicated in Table 1, the comparative analysis of the AI and human efficiencies in scoring EFL essays reveals that there are proportional deviations between the scores of ChatGPT and human raters. In contrast, the comparative efficiency indicates a proportional proximity or alignment between the scores of ChatGPT Edu and those of human raters. The gap between ChatGPT and human raters' scores is an average of 2.05, while that between ChatGPT Edu and human raters' scores is an average of 0.17.

ChatGPT Edu tends to predict the closer scores that are provided by actual human teachers when it comes to automatic essay scoring. Conversely, the scores from ChatGPT 3.5 are 30% lower than those from human raters (see Table 1). This suggests that ChatGPT Edu outperforms ChatGPT 3.5 in scoring essays, while ChatGPT Edu's scores are on par with those of human educators.

Table 2 : Comparison of the Scores Given by ChatGPT 3.5, ChatGPT Edu, and Human Teacher

Essay	Human Score	ChatGPT-3.5 Score	ChatGPT Edu Score	Difference (ChatGPT-Human)	Difference (ChatGPT Edu-Human)
1	8.5	6	8	2.5	0.5
2	9	5	9	4	0
3	8	6	8	2	0
4	7.5	6	7.5	1.5	0
5	7.5	5.5	8.5	2	1
6	7	5	7	2	0
7	6.5	4.5	6.5	2	0
8	6.5	5	6.5	1.5	0
9	5	4	5	1	0
Average				2.05	0.17

Further, ChatGPT Edu is likely to demonstrate stronger performance in essay scoring compared to ChatGPT 3.5. The scoring approach of ChatGPT Edu is broader as it scores over the breakdown of six criteria, while the ChatGPT 3.5 scoring approach is simpler and focuses on just two dimensions. ChatGPT Edu's scoring method is designed to mimic the thoroughness and fairness expected in human grading processes, as human teachers score across a broad spectrum of multi-dimensions or criteria (See Table 3).

Table 3: Comparison of the Numbers and Criteria along which ChatGPT 3.5, ChatGPT Edu, and Human Teacher Scores

Number and Names of Criteria ChatGPT 3.5 Scores along	Number and Names of Criteria ChatGPT Edu Scores along	Number and Names of Criteria Human Teacher Scores along
2 (Content and Structure, and language, and grammar	6 (content, language and vocabulary, organization and structure, grammar and syntax, cohesion and coherence, and overall impression	6 (Title, thesis statement, evidence, references, cohesion and coherence, and conclusion)

In addition, ChatGPT Edu's scores have more potential for consistent focus and interpretation than those of ChatGPT 3.5 because the ChatGPT Edu criteria combination is skill homogeneity-based, such as language and vocabulary, cohesion and coherence, organization and structure. In contrast, the ChatGPT 3.5 criteria combination is skill heterogeneity, such as content and structure.

On top of that, ChatGPT Edu's heterogeneous skill focus in its scoring approach tends to result in more comprehensive feedback comments than those of ChatGPT 3.5, because ChatGPT Edu feedback points to multiple areas by making far more comments than ChatGPT 3.5 (See Table 3). The human teacher feedback approach is much closer to ChatGPT Edu than it is to ChatGPT 3.5, as the teacher puts in many correction codes on various sub-skills while plowing through the essay.

Table 4 : Multi-Comment Feedback on Diverse Subskills of Writing

ChatGPT 3.5 Feedback Comments and Skill Areas	ChatGPT Edu Feedback Comments and Skill Areas	Human Teacher Feedback Comments and Skill Areas
Content and Structure: 5	Content and Structure: 9	Content and Structure: so many correction codes
Grammar, Syntax, Vocabulary, and language: 8	Language and Grammar: 4	Language and grammar: So many correction codes

B. Discussion

Since its inception in late 2022, Generative AI tools (GenAI), such as ChatGPT, have been speculated to bring forth substantial technological spillovers in education (Bahroun, Anane & Ahmedet al., 2023; Yusuf, Pervin & Román-González, 2024). Since its emergence, it has been re-versioned and tailor-made, resulting in ChatGPT 3.5, ChatGPT 40, and ChatGPT Edu. Therefore, more speculations are due around the comparative educational good that the various versions of ChatGPT can bring forth, their non-excludability in education, their techno-spillovers on educational systems, ecology, and determinism, as suggested by studies (Abramski et al., 2023; Koubaa, 2023; Scheschenja et al., 2024). Existing research suggests that there may be an asymmetry between the performance metrics of the earlier and newer versions of ChatGPT. The present study, thus, explores the revolutionary tech-positivity that ChatGPT Edu (ChatGPT's latest version tailor-made for education) can bring to education, more specifically, in EFL students' essay scoring.

Based on the findings in the above section, the present study acknowledges the nomenclature of ChatGPT as an "auto-grader" attributed by other scholars (Chang & Ginter, 2024), and similarly considers ChatGPT Edu as "an advanced auto-grader". The present paper discusses three premises to rationalize this attribution to and characterization of ChatGPT Edu: accuracy, fairness, and benchmarking.

In scoring the essays written by tertiary EFL students, ChatGPT Edu's comparative advantage is much higher than its predecessor ChatGPT 3.5. ChatGPT Edu's instrumental optimality is partly enhanced by its upgraded scoring precision. Traditionally, human scorers' standards are considered hermeneutic, and their scores are interpreted as quality essay scoring. As against this grand-value of human scoring, ChatGPT Edu is a more closely accuracy-aligned AI scorer (0.17) than ChatGPT 3.5 (2.05). These functionality gaps between the AI versions were signalled by previous studies as well (Jukiewicz, 2024). Further, previous studies (Javaid, Haleem & Singh, 2023; Jukiewicz, 2024) and the present study agree that ChatGPT scoring takes a turn towards a normativity where AI scoring accuracy increases with the advancements and plurality of AI tools. In line with that, because ChatGPT Edu carries more corresponding semblance to human scorers, a post-human approach may be taken to ChatGPT Edu for the co-evolution of humans and AI, and along that evolutionary trajectory, ChatGPT Edu may be pragmatized as more purpose-built for educational services-more specifically in essay scoring. Thus, ChatGPT Edu may be adopted by ELT practitioners for writing assessments to reduce their workload.

ChatGPT Edu's assessment is characterized by a methodological holism that accounts for fairness, objectivity, and justice. ChatGPT Edu distributes its scoring across criteria (6 criteria, twice as many as ChatGPT 3.5 (3 criteria). In line with Jukiewicz (2024), the present study finds that ChatGPT 3.5 scores essays across three criteria. This method accounts for a reductionist approach to essay assessment. In contrast, ChatGPT Edu's writing evaluation is less vulnerable to overgeneralization error, preventing epistemic injustice and unfairness due to its holistic approach to macro procedural distributiveness.

Furthermore, the present study finds that ChatGPT Edu passes more successfully than ChatGPT 3.5 in the Turing Test in terms of mimicking human teachers' benchmarking of writing pedagogy and assessment. ChatGPT Edu's scoring criteria outnumber those of ChatGPT 3.5, and therefore, the corrective feedback of the former also outnumbers that of the latter. Previous studies (Zhang & Wang, 2023; Brown, Smith & Lee, 2024) reported that ChatGPT achieves closeness to human teachers' essay scoring. The present study reports that ChatGPT Edu has brought about a higher degree of anthropomorphic alignment with human benchmarking by its mimesis. ChatGPT Edu's feedback is more dialogical, ensured by its broader framework of giving corrective feedback, compared to ChatGPT's narrow framework. The broader strategic communicative approach of ChatGPT Edu to assessing writing signals a more sociologically normative pedagogy that is constitutive of a three-party engagement: students' writing, ChatGPT Edu's higher degree of anthropomorphicity, and human teachers' necessitated monitoring. Precisely, the hybridity of ChatGPT Edu with a higher degree of proximity to human pedagogical benchmarking and traditions has the potential for intelligent and optimal student engagement in writing output and the zone of proximal development.

III. CONCLUSION

A. Conclusion

There is a reciprocal and cyclical relationship between the co-evolution of artificial intelligence (AI) and education. The *advancement* of AI continuously brings forth *new possibilities* in education. Given this universal normativity of AI spillovers on education, the emergence of ChatGPT and its optimization promise the potential to revolutionize educational practices. Hence, the present study aims to explore the comparative advantages of ChatGPT 3.5 and ChatGPT Edu in terms of the degree of correspondence of accuracy, fairness, and benchmarking to human teachers. Through a case-study method, the article concludes that compared to its earlier version, i.e., ChatGPT 3.5, ChatGPT Edu is potentially more instrumental in assessing EFL writing, more specifically, essay scoring, because of its superior accuracy, fairness, and benchmarking abilities. Because of its methodological holism—scoring across a broader set of criteria— ChatGPT Edu stands out as a more reliable

and equitable GenAI tool for EFL writing assessment. In addition, its dialogically comprehensive feedback framework more closely mimics the benchmarking of human teachers' writing and assessment. Its anthropomorphic composites signal a potential paradigm shift towards a co-evolutionary model between AI and human language pedagogues in which ChatGPT Edu could ease teacher workloads by facilitating more effective, fair, and comprehensive writing evaluation. Thus, the present study contributes to the understanding of ChatGPT Edu's promise for ELT practitioners, especially for writing pedagogues. The study also contributes to advancing the broader discourse on AI in ELT, and narrowly on the phenomenon of GPT's critical importance in terms of accuracy, fairness, and benchmarking in L2 writing assessments. It is also of paramount importance to note that this paper takes a secular and objective perspective in the analysis of the performance of the AI tools under investigation. The researchers do not take any subjective stimuli to exaggerate the positivity of ChatGPT Edu, nor do they underestimate the performance of ChatGPT 3.5. Moreover, the researchers believe that neither ChatGPT Edu nor ChatGPT 3.5 should be considered an educational panacea, and their real values depend on the way they are incorporated into educational contexts and practices.

B. Limitations and Scopes for Further Research

The study was undertaken as a rapid-response initiative following the emergence of Edu-GPT. For capturing the early insights into its performance, some methodological compromises were made, such as limitations in temporal rigour and sampling robustness. These limitations may affect the generalizability of the findings. To address these limitations and to help establish more reliable and generalizable empirical evidence regarding the AI tool's effectiveness in educational assessment contexts, further research involving larger and more diverse sample groups and utilizing advanced versions of the AI assessment tools is recommended.

C. Author Contributions

The contributions to this study were distributed among the authors as follows: Conceptualization**:** Md. Saiful Alam and A H M Ohidujjaman; Methodology**:** Md. Saiful Alam; Formal analysis**:** Md. Saiful Alam and A H M Ohidujjaman; Investigation**:** Md. Saiful Alam and A H M Ohidujjaman; Resources**:** Md. Saiful Alam, A H M Ohidujjaman, and Md. Nurullah Patwary; Data curation**:** Md. Saiful Alam and A H M Ohidujjaman; Writing—original draft preparation**:** Md. Saiful Alam; Writing—review and editing**:** Md. Nurullah Patwary and Humayra Akhter. All authors have read and agreed to the published version of the manuscript.

D. Funding

This work received no external funding. All resources used in its completion were provided solely by the author.

E. Acknowledgment

The principal author sincerely acknowledges the contributions of the second, third, and fourth authors of the article, who contributed enormously to the data collection, data analysis, and overall linguistic refinement of the study. The authors would also like to express their sincere gratitude to their colleagues for their all-out support in data collection and insightful comments and suggestions, which greatly enhanced the depth and quality of the article. Additionally, they extend their heartfelt gratitude to the scholarly peer reviewers and the dedicated editorial team members for their time, guidance, and critical input that were necessary to bring the article to its present standard.

F. Informed Consent

The authors have obtained informed consent from all participants.

G. Conflicts of Interest

The authors declare that there is no conflict of interest among them.

H. Data Availability

The data supporting the findings of the study are not publicly available due to privacy and ethical restrictions. However, the data may be available from the corresponding author upon reasonable request.

IV. REFERENCES

[1] Abramski Katherine, Citraro Salvatore, Lombardi Luigi, et al., Cognitive network science reveals bias in GPT-3, GPT-3.5 Turbo, and GPT-4 mirroring math anxiety in high-school students, Big Data and Cognitive Computing. 7(3) (2023) 124. https://doi.org/10.3390/bdcc7030124.

[2] Adeshola Ibrahim and Adepoju Adeola Praise, The opportunities and challenges of ChatGPT in education, Interactive Learning Environments. 32(10) (2023) 1-14. https://doi.org/10.1080/10494820.2023.2253858.

[3] Ajay Helen B., Tillett Paul I., and Page Ellis Betten, The analysis of essays by computer (AEC-II): Final report (Technical Report No. 8-0102), U. S. Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development. (1973)

[4] Altamimi Ahmed B., Effectiveness of ChatGPT in essay auto-grading. In: Proceedings of the 2023 International Conference on

Computing, Electronics & Communications Engineering (iCCECE); Aug 14-16, 2023; Swansea, UK. (2023)102–106. https://doi.org/10.1109/iCCECE59400.2023.10238541.

[5] Ali Kamran, Barhom Noha, Tamimi Faleh and Duggal, Monty, ChatGPT—A double-edged sword for healthcare education? Implications for assessments of dental students, European Journal of Dental Education. 28(1) (2023) 206-211. https://doi.org/10.1111/eje.12937

[6] Amineh Roya Jafari, and Asl Hanieh Davatgari, Review of constructivism and social constructivism, Journal of social sciences, literature and languages. 1(1) (2015) 9-16. https://www.blue-ap.com/J/List/4/iss/volume%2001%20(2015)/issue%2001/2.pdf

[7] Asmawi Adelina and Alam M. Saiful, Qualitative research: Understanding its underlying philosophies. Forum for Education Studies. 2 (2) (2024) 1320.https://doi.org/10.59400/fes.v2i2.1320

[8] Bahroun Zineb, Anane Chakib, Ahmed Véronique et al., The potential of ChatGPT in education: A systematic bibliometric review, Sustainability. 15(17) (2023) 12983. https://doi.org/10.3390/su151712983

[9] Božić Velibor and Poola, Indrasen, ChatGPT and education [preprint]. (2023) [cited 2025 Aug 13]; Available from: https://doi.org/10.13140/RG.2.2.18837.40168

[10] Bouziane Karima and Bouziane Abdelmounim, AI versus human effectiveness in essay evaluation, Discover Education. 3(1) (2024) 201. https://doi.org/10.1007/s44217-024-00320-6

[11] Brown Terry, Smith Rebecca and Lee Alan, How close is ChatGPT to human graders in assessing student essays?, Int J Artif Intell Educ. 34(1) (2024) 112-128. https://doi.org/10.1007/s40593-023-00312-9

[12] Bukowski Marcin and Tokowicz Nora, Automated essay scoring: A review of the literature, Behavior Research Methods. 53 (2021) 2090–2113. https://doi.org/10.3758/s13428-020-01509-1.

[13] Chang Li-Hsin and Ginter Filip, Automatic short answer grading for Finnish with ChatGPT. In: Proceedings of the AAAI Conference on Artificial Intelligence. 38(21), (Mar 2024) 23173–23181. Menlo Park (CA): AAAI Press (2024). https://doi.org/10.1609/aaai.v38i21.30363.

[14] Chen Shen, Li Yingya, Lu Sheng, et al., Evaluating the ChatGPT family of models for biomedical reasoning and classification, Journal of the American Medical Informatics Association. 31(4) (2024) 940-948. https://doi.org/10.1093/jamia/ocad256.

[15] Crompton Helen and Burke Diane, Artificial intelligence in higher education: the state of the field, International Journal of Educational Technology in Higher Education. 20(22) (2023) 1-22. https://doi.org/10.1186/s41239-023-00392-8

[16] Devitt Michael, Realism and Truth. Princeton, NJ, USA: Princeton University Press. (1997)

[17] deWinter Joost C.F., Dodou Dimitra and Stienen Arno H.A., ChatGPT in Education: Empowering Educators through Methods for Recognition and Assessment, Informatics. 10(4) (2023) 87. https://doi.org/10.3390/informatics10040087.

[18] Dong Fei, Zhang Yue and Yang Jie, Attention-based recurrent convolutional neural network for automatic essay scoring. In: Levy, R., Specia, L. editors. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017); 2017 Aug (3-4); Vancouver, Canada. Association for Computational Linguistics. (2024) 153–162. https://doi.org/10.18653/v1/K17-1017

[19] Dwivedi Yogesh, K., Kshetri Nir., Hughes Laurie, et al., "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges, and implications of generative conversational AI for research, practice, and policy, International Journal of Information Management. 71 (2023) 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642.

[20] Felix Catherine V., Ed., The role of the teacher and AI in education. In International perspectives on the role of technology in humanizing higher education. West Yorkshire, United Kingdom: Emerald Publishing Limited. (2020) 33-48.

[21] Firat Mehmet, How ChatGPT can transform autodidactic experiences and open education? [preprint], Center for Open Science. (2023). [cited 2025 Aug 4]. Available from: https://doi.org/10.31219/osf.io/9ge8m

[22] Flyvbjerg Bent, (Ed.), Case study. In: Denzin NK, Lincoln YS, editors. The Sage Handbook of Qualitative Research. (4th ed.). Thousand Oaks (CA), USA: Sage Publications. (2011) 301–316.

[23] Foltz Peter W., Laham David and Landauer Thomas K., Automatic essay assessment, Assessment in Education: Principles, Policy & Practice. 10(3) (2003) 295–308. https://doi.org/10.1080/0969594032000148154.

[24] Gardner John, O'Leary Michael, and Yuan Li, Artificial intelligence in educational assessment: 'Breakthrough? or buncombe and ballyhoo?, Journal of Computer Assisted Learning. 37(5) (2021) 1207-1216. https://doi.org/10.1111/jcal.12577

[25] Harré Romano and Krausz Michael, Varieties of Relativism. Oxford, UK: Blackwell. (1996)

[26] Herbold Steffen, Hautli-Janisz Annette, Heuer Utte, et al., A large-scale comparison of human-written versus ChatGPT-generated essays, Sci. Rep. 13(1) (2023)18617. https://doi.org/10.1038/s41598-023-45644-9.

[27] Jackaria Potching M., Hajan Bonjovi H., and Mastul Al-Rashiff H., A comparative analysis of the rating of college students' essays by ChatGPT versus human raters, International Journal of Learning, Teaching and Educational Research. 23(2) (2024) 478-492. https://doi.org/10.26803/ijlter.23.2.23

[28] Javaid Mohd, Haleem Abid and Singh Ravi Pratap et al., Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system, BenchCouncil Transactions on Benchmarks, Standards and Evaluations. 3(2) (2023) 100115. https://doi.org/10.1016/j.tbench.2023.100115

[29] Jukiewicz Marcin, The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process, Thinking Skills and Creativity. 52 (2024) 101522. https://doi.org/10.1016/j.tsc.2024.101522

[30] Kayyali Mustafa, Ed., Future possibilities and challenges of AI in education. In: Sharma RC, Bozkurt A, editors, Transforming education with generative AI: prompt engineering and synthetic content creation. 1st ed. Pennsylvania, United States, Hershey: IGI Global. (2024) 118–37.

[31] Kınık Busra and Çetin Husein, Human vs. AI: The use of ChatGPT in writing assessment, Advances in Educational Technologies and Instructional Design Book Series. Hershey, Pennsylvania (PA), United States: IGI Global. (2023) 194–215. https://doi.org/10.4018/979-8-3693-0353-5.ch009

[32] Koubaa Anis, GPT-4 vs. GPT-3.5: A concise showdown [preprint], Preprints.org; 2023 Mar 24 [cited 2025 Aug 4]. Available from: https://doi.org/10.20944/preprints202303.0422.v1
[33] Lampropoulos Georgios and Papadakis Stamatios, (Ed.), The Educational Value of Artificial Intelligence and Social Robots. In: Lampropoulos, G., Papadakis, S., Social Robots in Education. Studies in Computational Intelligence. Cham, Canton of Zug, Switzerland: Springer. (2025) 3-15.
[34] Latif Ehan and Zhai Xiaoming, Fine-tuning ChatGPT for automatic scoring, Computers and Education: Artificial Intelligence. 6 (2024) 100210. https://doi.org/10.1016/j.caeai.2024.100210
[35] Michel-Villarreal Rosario, Vilalta-Perdomo Eliseo, Salinas-Navarro David Ernesto, et al., Challenges and opportunities of generative AI for higher education as explained by ChatGPT. Education Sciences. 13(9) (2023) 856. https://doi.org/10.3390/educsci13090856
[36] Mizumoto Atushi and Eguchi Masaki, Exploring the potential of using an AI language model for automated essay scoring, Research Methods in Applied Linguistics. 2(2) (2023) 100050. https://doi.org/10.1016/j.rmal.2023.100050.
[37] Okunlaya Rifqah Olufunmilayo, Syed Abdullah Norris and Alinda Alias Rose, Artificial intelligence (AI) library services innovative conceptual framework for the digital transformation of university education, Library Hi Tech. 40(6) (2022) 1869-1892. https://doi.org/10.1108/LHT-07-2021-0242
[38] Page Ellis Betten, The use of the computer in analyzing student essays, International Review of Education. 14(2) (1968) 210–225. http://www.jstor.org/stable/3442515.
[39] Page Ellis Betten, The imminence of... grading essays by computer, Phi Delta Kappan. 47(5) (1966) 238–243. Available from: https://www.jstor.org/stable/20371545
[40] Page Ellis Betten, Ed., Project Essay Grade: PEG. In Shermis, M. D., Jill B., Automated essay scoring: A cross-disciplinary perspective. Mahwah, New Jersey, USA: Lawrence Erlbaum Associates. (2003) 43-45.
[41] Polverini Guilia and Gregorcic Bor, How understanding large language models can inform the use of ChatGPT in physics education, European Journal of Physics. 45(2) (2025) 025701. https://:10.1088/1361-6404/ad1420
[42] Potchong, M., Hajan, Bonjovi H., and Mastul, Al-Rashiff H., A comparative analysis of the rating of college students' essays by ChatGPT versus human raters, Int'l Journal of Learning, Teaching and Educational Research. 23(2) (2024) 478–92. https://doi.org/10.26803/ijlter.23.2.23
[43] Ramesh Dadi and Sanampudi Suresh Kumar, An automated essay scoring system: a systematic literature review, Artificial Intelligence Review. 55(3) (2022) 2495-2527. https://doi.org/10.1007/s10462-021-10068-2
[44] Reddit community post, OpenAI introduces ChatGPT Edu for universities [Online forum post], Reddit. (May 30, 2024). https://www.reddit.com/r/OpenAI/comments/1d4df2e/
[45] Rospigliosi Pricles Asher, Artificial intelligence in teaching and learning: what questions should we ask of ChatGPT?, Interactive Learning Environments. 31(1) (2023) 1-3. https://doi.org/10.1080/10494820.2023.2180191
[46] Rudner Lawrence M., and Liang Tahung, Automated essay scoring using Bayes' theorem, The Journal of Technology, Learning and Assessment. 1(2) (2002). [cited 2025 Aug 13]; Available from: https://ejournals.bc.edu/index.php/jtla/article/view/1668
[47] Shafik Wasswaa, Introduction to ChatGPT. In Advanced Applications of Generative AI and Natural Language Processing Models. Hershey, Pennsylvania, United States: IGI Global. (2024) 1-25.
[48] Scheschenja Michael, Viniol Simon, Bastian Moritz B., et al., Feasibility of GPT-3 and GPT-4 for in-depth patient education prior to interventional radiological procedures: a comparative analysis, Cardiovascular and interventional radiology. 47(2) (2024) 245-250. https://doi.org/10.1007/s00270-023-03563-2.
[49] Schiff Daniel, 2022. Education for AI, not AI for education: The role of education and ethics in national AI policy strategies, International Journal of Artificial Intelligence in Education. 32(3) (2022) 527-563. https://doi.org/10.1007/s40593-021-00270-2
[50] Shermis Mark D., Mzumara Howard, Olson Jenifar, et al., On-line grading of student essays: PEG goes on the World Wide Web. Assessment & Evaluation in Higher Education. 26(3) (2001) 247–258. https://doi.org/10.1080/02602930120052404.
[51] Slinkard Karen and Singleton Vernon L., Total phenol analysis: automation and comparison with manual methods, American Journal of Enology and Viticulture. 28(1) (1997) 49-55. https://doi.org/10.5344/ajev.1974.28.1.49.
[52] Steiss Jacob, Tate Tamara, Graham Steve, et al., Comparing the quality of human and ChatGPT feedback of students' writing, Learning and Instruction. Jun 1 (2024) 91:101894–4. https://doi.org/10.1016/j.learninstruc.2024.101894
[53] Su Jiahong and Yang Weipemg, Unlocking the power of ChatGPT: A framework for applying generative AI in education, ECNU Review of Education. 6(3) (2023) 355-366. https://doi.org/10.1177/20965311231168423
[54] Tahiru Fati, AI in education: A systematic literature review, Journal of Cases on Information Technology (JCIT). 23(1) (2021) 1- 20. https://doi.org/10.4018/JCIT.2021010101
[55] Yusuf Abdullah, Pervin Nasrin and Román-González Marcos, Generative AI and the future of higher education: a threat to academic integrity or reformation? Evidence from multicultural perspectives, Int J Educ Technol High Educ. 21(1) (2024) 21. https://doi.org/10.1186/s41239-024-00453-6
[56] Zhang Liang and Wang Hao, Automated essay scoring using generative AI models: An empirical study with ChatGPT, Comput Educ. (2023) 180:104524. https://doi.org/10.1016/j.compedu.2022.104524.
[57] Zhai Xiaoming and Nehm Ross H., AI and formative assessment: The train has left the station, Journal of Research in Science Teaching. 60(6) (2023) 1390-1398. https://doi.org/10.1002/tea.21885
[58] Zhai Xuesong, Xiaoyan Chu, Ching Sing Chai, et al., A Review of Artificial Intelligence (AI) in Education from 2010 to 2020, Complexity. (1) (2021) 8812542. https://doi.org/10.1155/2021/8812542