THE ROLE OF ARTIFICIAL INTELLIGENCE IN ASSESSING ENGLISH LANGUAGE SKILLS: OPPORTUNITIES AND LIMITATIONS.

Maxmanazarova Feruza Ismoilovna
Senior teacher of UZSWLU
Ergasheva Marjonaxon Muhammadzokir qizi.
Student of UZSWLU

Abstract. *This study investigates the effectiveness of artificial intelligence (AI) in assessing English language writing skills by comparing AI-generated scores with human examiner evaluations. Employing a qualitative and comparative research design, the study analysed essays written by four intermediate (B1) EFL university students. Each essay was assessed by both an AI-based scoring system and experienced human raters using standard writing criteria. The results revealed a moderate level of agreement between the two scoring methods, with 40% of the scores being identical, 45% showing slight differences, and 15% demonstrating significant discrepancies. The findings indicate that while AI systems are capable of reliably evaluating surface-level linguistic features such as grammar and vocabulary, they remain limited in assessing higher-order aspects of writing, including coherence, argumentation, and communicative effectiveness. Additionally, AI tended to assign slightly lower scores than human examiners, reflecting a more rigid and rule-based approach. The study concludes that AI-based assessment can serve as a valuable supportive tool in language evaluation; however, it should complement rather than replace human judgment. A combined approach is recommended to ensure more accurate, fair, and meaningful assessment of writing proficiency.*

Keywords *- Artificial Intelligence (AI); Automated Essay Scoring; Language Assessment; Writing Skills; Human vs AI Evaluation; EFL Learners; Educational Technology (EdTech); Writing Proficiency; Assessment Reliability; Digital Assessment*

Introduction

The growing use of artificial intelligence in English language testing marks a clear shift away from paper-based assessment, creating both new opportunities and important limitations for evaluating language skills.

"Artificial intelligence is becoming a key force in the EdTech field because it helps companies automate routine tasks, analyse student data and provide fast feedback to learners" (George & Wooden, 2023). For example, platforms such as Squirrel AI use AI to offer personalised tutoring, identify learners' weaknesses and recommend suitable learning materials, while natural language processing and machine-learning methods are also used to develop intelligent tutoring systems that can interact with students in a more human-like way and improve the learning experience (Kökver et al., 2024). At the same time, collaborative learning is supported through online platforms such as Google Classroom and Microsoft Teams, which allow students to work together, share resources and give peer feedback. In addition, gamification is widely used to increase learners' motivation and engagement, and platforms such as Kahoot! and Classcraft apply game elements to traditional learning activities. Finally, EdTech companies rely on data analytics to collect and examine large amounts of information about students'

DOI: 10.5281/zenodo.19780608

performance, learning behaviour and engagement, which helps teachers make better decisions about curriculum design, resource use and targeted support.

Several studies show that many teachers remain cautious about the value of educational technologies. Nearly 60% of educators report doubts about whether technology can truly improve students' learning outcomes, and a large number still prefer traditional teaching approaches (Fishman et al., 2016; Ghavifekr & Rosdy, 2015; Israel et al., 2015). In addition, about 70% of teachers state that they have not received enough training to use digital tools effectively, which points to a strong need for EdTech companies to offer clear and well-structured professional development programmes (Bragg et al., 2021). Furthermore, many educators are uncertain about the long-term sustainability of technology-based initiatives, as rapid technological change and fears that tools may quickly become outdated reduce teachers' willingness to fully adopt new systems in their teaching practice (McDiarmid & Zhao, 2023; Miller, 2014).

Additionally, Transformative learning theory, first introduced by Jack Mezirow and later developed by Patricia Cranton and Edward Taylor, offers an important way to understand how education and technology influence each other by explaining how meaningful learning experiences can change people's beliefs and teaching practices. In the field of EdTech, the rapid move to digital platforms during the COVID-19 pandemic became a powerful transformative experience for both teachers and students, as it challenged traditional classroom methods and encouraged new ways of thinking about teaching and learning. Even before the pandemic, the EdTech sector was already developing quickly due to technological innovation, changes in teaching approaches and a growing demand for flexible and personalised learning, including the wider use of online learning platforms, the introduction of artificial intelligence and adaptive learning systems, and the growing use of virtual and augmented reality tools, all of which expanded opportunities for personalised and more immersive learning and gradually reshaped traditional educational practices.

Automated essay scoring systems have been widely criticized for placing significant emphasis on surface linguistic features, such as grammatical accuracy, sentence length, and lexical complexity, while often neglecting deeper elements of writing quality. Studies indicate that because these systems depend on patterns derived from training data, they may fail to recognize aspects such as creativity, originality, and sophisticated argumentation. As a result, AI-based assessments may provide incomplete or potentially biased evaluations of student writing, particularly for learners who come from diverse linguistic or cultural backgrounds. For this reason, many researchers suggest that automated scoring should be used alongside human evaluation in order to ensure a more balanced and reliable assessment of language proficiency.

One of the primary concerns surrounding the widespread use of EdTech is the collection and use of student data. EdTech platforms gather vast amounts of data on learners, including their learning preferences, progress and behaviour. While this data can be invaluable for personalising learning experiences and improving educational outcomes, it also raises significant privacy and security concerns.3 Ensuring the responsible use and protection of student data is crucial to maintaining trust and protecting learners from potential risks. While many developers claim their AI is accurate, there is a lack of research comparing these scores to the judgments of expert human examiners. This study aims to fill that gap.

Literature Review

The integration of artificial intelligence (AI) into education has significantly transformed language teaching and assessment practices. In recent years, educational technology (EdTech) has developed rapidly, providing new tools that support both learning and evaluation. According to George and Wooden, artificial intelligence has become an important driving force in the EdTech sector because it allows companies to automate routine processes, analyse large volumes of student data and deliver immediate feedback to learners. Such technological developments have made it possible to create more personalised and adaptive learning environments that respond to individual learner needs.

One of the major developments in AI-supported education is the use of intelligent tutoring systems and personalised learning platforms. For instance, the AI-based learning platform Squirrel AI uses machine learning algorithms to analyse students' performance, identify weaknesses and recommend appropriate learning materials. Research by Kökver and colleagues highlights that natural language processing and machine learning techniques are increasingly used to design intelligent tutoring systems capable of interacting with learners in ways that resemble human communication. These systems can provide personalised feedback, adjust task difficulty and support learners in developing their language skills more effectively.

In addition to AI-driven systems, digital platforms have also contributed to the development of collaborative learning environments. Platforms such as Google Classroom and Microsoft Teams allow students to communicate, exchange resources and engage in peer feedback activities. These tools encourage interaction and cooperation among learners while supporting teachers in managing learning activities in online or hybrid classrooms. Another important development in EdTech is the use of gamification to increase student motivation and engagement. Learning platforms such as Kahoot! and Classcraft integrate game-like elements into educational activities, which can make learning more interactive and enjoyable for students.

Despite these advantages, many educators remain cautious about the effectiveness of technology-based learning tools. Research conducted by Fishman and colleagues suggests that a significant proportion of teachers question whether educational technologies truly improve learning outcomes. Similarly, studies by Ghavifekr and Rosdy indicate that many teachers still prefer traditional teaching approaches. One of the main reasons for this hesitation is the lack of sufficient training in the use of digital technologies. According to research by Bragg and colleagues, a large number of teachers report that they have not received adequate professional development to effectively integrate digital tools into their teaching practice. In addition, rapid technological changes and concerns about the sustainability of technology-based initiatives may discourage teachers from fully adopting new systems in education (McDiarmid & Zhao, 2023; Miller, 2014).

The relationship between technology and educational change can also be explained through the perspective of Transformative Learning Theory, originally introduced by Jack Mezirow and further developed by Patricia Cranton and Edward Taylor. This theory suggests that significant learning experiences can transform individuals' beliefs and perspectives, leading to changes in their professional practices. In the context of education technology, the rapid shift to digital learning during the COVID-19 pandemic served as a transformative experience for both educators and learners. The sudden move

to online platforms challenged traditional teaching methods and encouraged the adoption of new digital tools and instructional strategies.

Although AI technologies offer many benefits in language education, researchers have also identified several important limitations. One of the most widely discussed concerns relates to automated essay scoring systems. These systems often focus heavily on surface linguistic features such as grammar accuracy, sentence length and vocabulary complexity, while overlooking deeper aspects of writing quality. As a result, they may fail to recognise elements such as creativity, originality and complex argumentation. Because automated scoring models rely on patterns learned from training data, they may produce incomplete or biased evaluations of student writing, particularly for learners from diverse linguistic or cultural backgrounds. For this reason, many scholars argue that AI-based scoring should complement rather than replace human assessment in order to ensure a more accurate and reliable evaluation of language proficiency.

Another critical issue related to the use of AI in education concerns data privacy and security. EdTech platforms collect large amounts of information about learners, including their learning preferences, progress and behavioural patterns. While such data can be used to improve personalised learning and educational outcomes, it also raises important ethical concerns about the protection and responsible use of student information. Ensuring transparency, security and ethical data management is therefore essential to maintaining trust in AI-based educational systems.

Overall, existing research highlights both the opportunities and challenges associated with the use of artificial intelligence in language education and assessment. While AI technologies provide powerful tools for personalisation, feedback and large-scale assessment, concerns remain regarding validity, fairness, transparency and data privacy. Moreover, there is still limited research comparing AI-generated scores with the evaluations of expert human examiners. Addressing this gap is essential for understanding the reliability of AI-based assessment systems and for determining how they can be effectively integrated into English language testing practices.

Methods

1 Research Design

This study employed a **qualitative and comparative research design** to investigate the effectiveness of artificial intelligence (AI) in assessing English language writing skills. The qualitative approach was used to analyse patterns and differences in scoring, while the comparative design allowed for a direct comparison between AI-generated scores and human examiner evaluations. This combination made it possible to identify similarities, inconsistencies, and potential limitations of AI-based assessment systems.

2 Participants

The participants of this study consisted of 4 **university students** studying English as a Foreign Language (EFL) at an **intermediate (B1) level**. The participants were selected based on their similar language proficiency to ensure consistency in the data. All students had prior experience in writing short argumentative essays in English.

3 Data Collection

Data were collected through a **writing task**. Each participant was asked to produce a **short essay** on different topics. The essays were designed to reflect typical academic writing tasks and to assess students' ability to organise ideas, use appropriate vocabulary, and demonstrate grammatical accuracy.

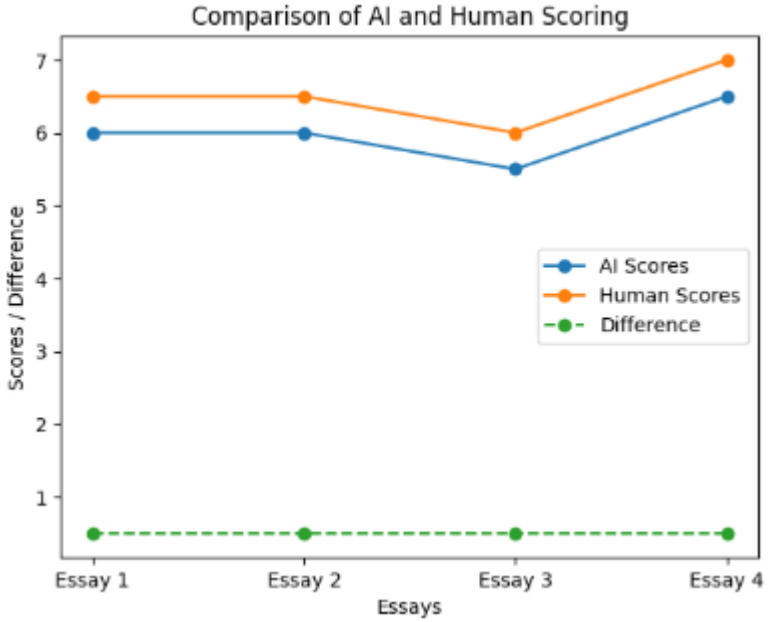
After collection, each essay was evaluated in two ways:

- First, using an **AI-based scoring system**, which automatically assessed the essays based mostly on linguistic features such as grammar, vocabulary, coherence, cohesion and structure.
- Second, by **human examiners**, who were experienced English language teachers. They evaluated the essays using mostly standard assessment criteria, including coherence, cohesion, task response, lexical resource, and grammatical accuracy.

4 Data Analysis

The data analysis in this study centered on a systematic comparison between scores generated by the AI-based assessment system and those assigned by human examiners. The primary objective of this comparison was to evaluate the level of agreement between the two scoring approaches and to identify any discrepancies in their evaluations. Specifically, the analysis examined (a) the extent to which AI-generated scores aligned with human ratings, (b) the degree of variation in scoring, and (c) emerging patterns in cases of disagreement.

Figure 1.



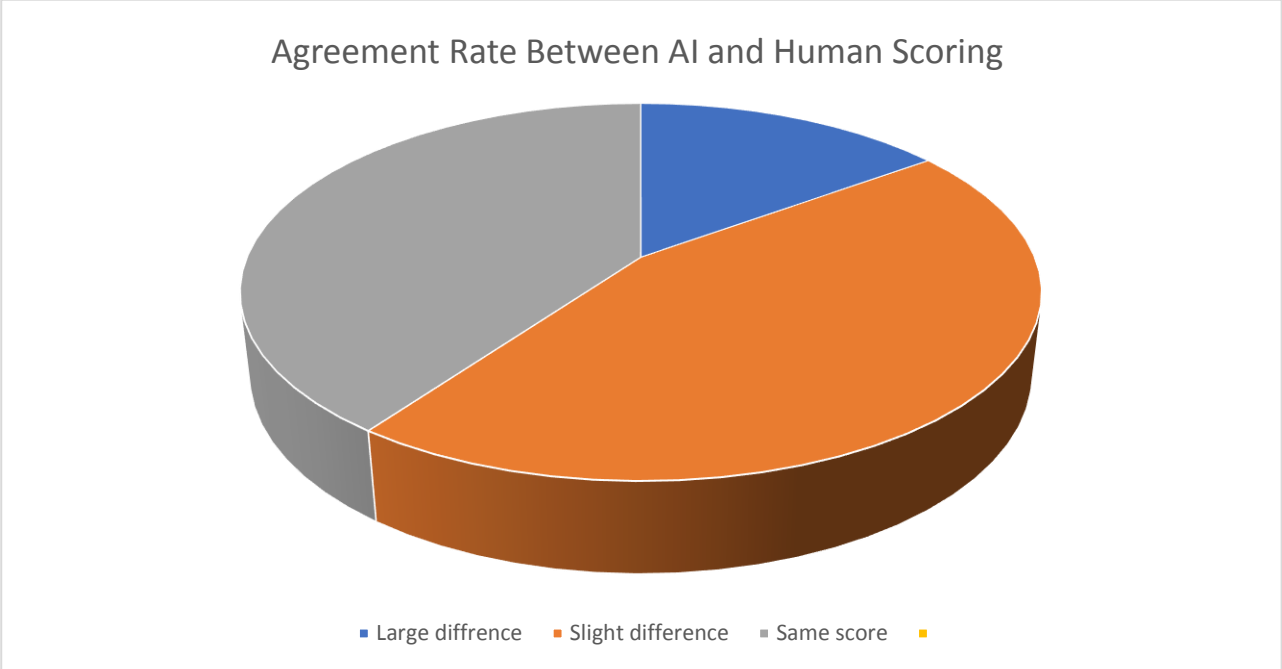
A sample of four essays was analyzed, with results indicating minor differences between AI and human scores. For instance, Essay 1 and Essay 2 both received a score of 6.0 from the AI system and 6.5 from human examiners, while Essay 3 was rated 5.5 by AI and 6.0 by humans. Essay 4 demonstrated a slightly higher evaluation, with AI assigning a score of 6.5 compared to the human score of 7.0. These findings suggest a consistent tendency for the AI system to assign slightly lower scores than human raters.

To provide a clearer interpretation of agreement levels, the results were categorized into three distinct groups: (1) identical scores, where AI and human evaluations fully matched; (2) slight differences, where the discrepancy between scores was minimal; and (3) large differences, where a significant gap was observed. Following this classification, the frequency of each category was calculated and converted into percentages to illustrate the overall agreement rate between AI and human scoring. This categorization enabled a more structured understanding of the reliability and consistency

of AI-based assessment in comparison to traditional human evaluation.

Results

Figure 2



After analysing the data, the results showed varying levels of agreement between the AI-generated scores and the human examiner evaluations. The findings are presented in Figure 2, which illustrates the agreement rate between the two scoring methods.

Overall, the comparison revealed that **40% of the essays received the same score** from both the AI system and human examiners. This indicates a moderate level of consistency between the two assessment methods.

In addition, **45% of the essays showed slight differences** in scoring. These differences were minor and did not significantly affect the overall evaluation of students' writing performance. This suggests that while AI assessment is generally close to human judgement, small variations still occur. However, **15% of the essays demonstrated large differences** between AI and human scores. In these cases, the AI system either overestimated or underestimated the quality of writing compared to human examiners. This highlights potential limitations in AI-based assessment, particularly in evaluating more complex aspects of writing. In summary, the results indicate that although there is a reasonable level of agreement between AI and human scoring, discrepancies still exist, especially in more nuanced evaluations of writing quality.Furthermore, the analysis revealed several notable differences between AI-generated scores and those assigned by human examiners. First, a consistent pattern was observed in scoring severity, with human raters tending to assign slightly higher scores than the AI system across all evaluated essays. This indicates that the AI system applies stricter assessment criteria, whereas human examiners demonstrate a greater degree of leniency. Second, differences emerged in the focus of evaluation: human feedback primarily emphasized communicative effectiveness, lexical appropriacy, and the natural

use of language, while AI feedback concentrated on measurable linguistic features such as grammatical accuracy, vocabulary classification, and structural elements. Furthermore, the types of errors identified varied considerably. Human raters were more attentive to context-dependent issues, including awkward phrasing, collocational inaccuracies, and coherence-related problems, whereas the AI system predominantly detected surface-level errors such as grammatical mistakes, spelling issues, and sentence-level inconsistencies. In addition, variations were evident in feedback style, with human evaluations providing more explanatory and example-based comments, while AI feedback was more systematic and categorical in nature. Finally, in higher-quality essays, human examiners demonstrated a greater focus on argumentation, coherence, and task achievement, whereas the AI system emphasized formal structural requirements. Overall, these findings highlight a fundamental distinction between holistic, meaning-oriented human assessment and rule-based, analytical AI evaluation.

Figure 3

Discussion

The present study set out to investigate the extent to which AI-based scoring aligns with human evaluation in the assessment of English language writing. The findings reveal a moderate level of agreement between the two scoring approaches, thereby contributing to the growing body of research on the role of artificial intelligence in language assessment. Overall, the results indicate that AI systems are capable of producing scores that are broadly comparable to those assigned by human examiners, particularly in cases where writing quality can be evaluated through clearly measurable linguistic features. However, the analysis also highlights several important discrepancies that point to fundamental differences in how AI and human raters interpret and assess written language.

One of the most significant findings of this study is the relatively high proportion of agreement between AI and human scoring, with 40% of essays receiving identical scores and an additional 45% showing only slight differences. This suggests that AI systems demonstrate a reasonable level of reliability in evaluating writing performance, especially at an intermediate (B1) proficiency level where linguistic features such as grammar, vocabulary, and sentence structure play a central role. These findings support earlier research indicating that automated essay scoring systems are particularly effective in assessing surface-level aspects of language, as they rely on predefined algorithms and large datasets to identify patterns of correctness and error. From a practical perspective, this level of agreement suggests that AI could serve as a useful tool in reducing the workload of human examiners and increasing the efficiency of large-scale language assessments.

Despite this overall consistency, the presence of discrepancies—especially the 15% of cases with large differences—raises important concerns about the limitations of AI-based evaluation. These differences are not merely statistical variations but reflect deeper issues related to the nature of writing assessment itself. Writing is a complex, multidimensional skill that involves not only linguistic accuracy but also the ability to express ideas clearly, construct logical arguments, and engage the reader effectively. While AI systems are well-equipped to evaluate quantifiable features such as grammar and lexical range, they often struggle to assess more abstract qualities such as coherence, cohesion, creativity, and communicative effectiveness. The findings of this study clearly demonstrate that human examiners are more sensitive to these higher-order aspects of writing, which may explain why discrepancies occur in cases where deeper interpretation is required.

Another notable pattern identified in the results is the consistent tendency of the AI system to assign slightly lower scores compared to human raters. This pattern suggests that AI operates with stricter and more rigid assessment criteria, likely due to its reliance on rule-based algorithms and statistical models. Unlike human examiners, who can interpret meaning flexibly and consider the overall effectiveness of communication, AI systems tend to penalise deviations from standard linguistic norms more heavily. This finding aligns with previous studies that argue automated scoring systems may disadvantage learners who use unconventional or creative language, even when their writing is communicatively successful. Consequently, the stricter nature of AI scoring may raise concerns about fairness, particularly for learners from diverse linguistic and cultural backgrounds.

In addition to differences in scoring patterns, the study also revealed clear distinctions in the focus of evaluation between AI and human raters. Human examiners were found to prioritise communicative effectiveness, lexical appropriacy, and the natural flow of language. Their evaluations often reflected an understanding of the writer's intention and the overall clarity of the message. In contrast, AI feedback was primarily oriented towards measurable linguistic features, including grammatical accuracy, vocabulary classification, and structural organisation. While this analytical approach allows for consistency and objectivity, it may result in a narrower interpretation of writing quality. This difference in evaluative focus highlights a key limitation of current AI systems, which tend to assess writing as a collection of discrete features rather than as an integrated act of communication.

Furthermore, the analysis of feedback styles provides additional insight into the qualitative differences between AI and human assessment. Human feedback was generally more detailed, explanatory, and context-sensitive, often including specific examples and suggestions for improvement. This type of feedback is particularly valuable for learners, as it not only identifies errors but also supports the development of writing skills through meaningful guidance. In contrast, AI-generated feedback tended to be more structured and categorical, frequently labelling errors without fully explaining their context or impact on overall communication. While such feedback may be efficient and easy to process, it may not always support deeper learning or help students understand how to improve their writing in a meaningful way.

These findings have important implications for the integration of AI into language assessment practices. First and foremost, the results support the argument that AI should be used as a complementary tool rather than a replacement for human evaluation. While AI offers clear advantages in terms of speed, consistency, and scalability, it lacks the interpretive and contextual sensitivity that human examiners bring to the assessment process. A hybrid approach, in which AI is used for initial scoring or diagnostic feedback and human raters provide final evaluation and qualitative insights, may offer the most effective solution. Such an approach would combine the strengths of both systems while minimising their respective limitations.

In addition, the findings highlight the need for ongoing development and refinement of AI-based assessment systems. In particular, there is a need to improve the ability of these systems to evaluate higher-order writing skills, including coherence, argumentation, and creativity. Advances in natural language processing and machine learning may help address these challenges by enabling AI systems to better understand context, meaning, and discourse-level features. However, achieving this level of sophistication remains a complex task, and further research is required to ensure that AI-based assessments are both valid and fair.

It is also important to consider the limitations of the present study. The small sample size, consisting of only four participants, limits the generalisability of the findings. While the study provides valuable insights into the comparison between AI and human scoring, future research should involve a larger and more diverse sample in order to produce more robust and generalisable results. Additionally, the study focused on a specific proficiency level (B1), and it is possible that the level of agreement between AI and human scoring may vary across different proficiency levels or types of writing tasks. Future studies could explore these variables in greater depth to provide a more comprehensive understanding of AI-based assessment.

Finally, the findings of this study can be interpreted within the broader context of Transformative Learning Theory. The increasing use of AI in education represents not only a technological shift but also a transformation in how assessment is conceptualised and implemented. As educators adapt to new tools and approaches, their beliefs and practices may also evolve, leading to new models of teaching and evaluation. However, this transformation must be approached critically, with careful consideration of both the opportunities and the challenges associated with AI. Ensuring that technological innovation enhances rather than undermines the quality and fairness of education remains a key priority.

In conclusion, this study demonstrates that AI-based scoring systems have the potential to play a significant role in language assessment, particularly in terms of

efficiency and consistency. However, important limitations remain, especially in relation to the evaluation of complex and nuanced aspects of writing. The differences observed between AI and human scoring highlight the continued importance of human judgment in the assessment process. Therefore, a balanced and integrated approach that combines the strengths of both AI and human evaluation is likely to be the most effective way forward in ensuring accurate, fair, and meaningful assessment of language proficiency.

References:

1. George, B., & Wooden, O. (2023). Managing the strategic transformation of higher education through artificial intelligence. *Administrative Sciences*, *13*(9), 196.
2. Kökver, Y., Pektaş, H. M., & Çelik, H. (2024). Artificial intelligence applications in education: Natural language processing in detecting misconceptions. *Education and Information Technologies*, 1-32. https://link.springer.com/article/10.1007/s10639-024-12919-1
3. Kim, H. (2019), 'The Perception of Teachers and Learners towards an Exploratory Corpus-based Grammar Instruction in a Korean EFL Primary School Context', *The Korea Association of Primary English Education*, Vol. 25, No. 1, pp. 123–152.
4. Christensen, C. (1997), *The Innovator's Dilemma*, Harvard Business School Press, Cambridge, MA.
5. AIGillian K. Hadfield 'Can AI Be Governed? Only If WeBuild Normatively Competent https://onlinelibrary.wiley.com/doi/epdf/10.1002/9781394258840.ch29