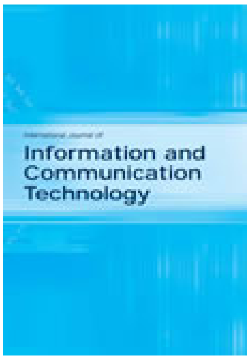
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Transforming English language education with AI-driven deep learning models for scalable adaptive and inclusive assessment

Yanwen Chen

DOI: 10.1504/IJICT.2025.10070759

Article History:

Received:	19 February 2025
Last revised:	01 March 2025
Accepted:	01 March 2025
Published online:	08 May 2025

Transforming English language education with AI-driven deep learning models for scalable adaptive and inclusive assessment

Yanwen Chen

Shanxi Institute of Technology,
Yangquan City, Shanxi Province, 045000, China
Email: chenyanwen122@163.com

Abstract: This study explores how deep learning can help English language education overcome the disadvantages of conventional means of assessment that are typically slow, subjective and not easily scalable. The research uses modern NLP models such as BERT and GPT to create AI-powered systems that assess reading, writing, listening, and speaking skills. They offer personalised real-time feedback designed specifically for any learner. A mixed methods approach combines educator and student insights with performance metrics. Results confirm that AI nearly doubles the accuracy, halves the grading time, and increases note engagement by 50% with clearly gained student proficiency. The study deals with the issue of cultural bias and privacy but uses it ethically and inclusively. By showing deep learning's great promise for providing fair, scalable and effective language learning, this multimodal framework represents an archetype they say can help overcome that challenge.

Keywords: deep learning models; natural language processing; NLP; English language education; AI-driven assessments; personalised feedback; multimodal language evaluation.

Reference to this paper should be made as follows: Chen, Y. (2025) 'Transforming English language education with AI-driven deep learning models for scalable adaptive and inclusive assessment', *Int. J. Information and Communication Technology*, Vol. 26, No. 12, pp.15–31.

Biographical notes: Yanwen Chen is a researcher at the Shanxi Institute of Technology (Yangquan City, Shanxi Province, China). She publishes academic research and practical application of technology, focusing on engineering. She is involved in regional scientific initiatives to advance innovation and education. Her work shows that one is committed to academic excellence and technological progress.

1 Introduction

In recent years, education technology has increased exponentially, rendering such roles fundamentally changing how learners access information, interact with content, and learn new skills (Hawamdeh and Abdelhafid, 2024; Hossain, 2023). In language education, new tools have emerged that help bypass traditional limitations in this shift. The evaluation techniques used for language acquisition are essential to identify what needs to be worked on, how to provide constructive feedback, and how to monitor learner progress. Traditional evaluation methods often fail to deliver consistent and scalable assessments. They require significant labour, suffer from inconsistencies, and are prone to human error (Nicol and Macfarlane-Dick, 2006; Hill and Miller, 2013; Sadler, 1989). The arrival of artificial intelligence (AI) and its technologies, deep learning, solves these problems as languages could be automated through grammar correction, pronunciation analysis, and contextual understanding.

Machine learning has already provided us with the possibility of AI-driven tools like Grammarly and Duolingo in language education (Beevi et al., 2024; Varun and Sathish, 2024; Čilić et al., 2024). However, as these tools personalise the learning experience by adapting to individual user needs, providing immediate feedback, and encouraging engagement, they appear to function for the instructor and student. These are the deep learning models: bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018) and generative pre-trained transformer (GPT) (Yenduri et al., 2024) at the front of this revolution. Indeed, their aptitude to process and humanly generate text has been accommodated in numerous applications, including automated essay grading and conversational practice (Sweta, 2024). The growing adoption of AI and deep learning signifies a significant shift in education, where technology plays a leading role in supporting educators and assisting students in their learning.

On the one hand, much progress has been made from a theoretical and methodology point of view, but traditional evaluation techniques in English language education remain the dominant technology (Wright, 2010). These methods are time-consuming and not consistent. Feedback quality from human evaluators is subjective and varies, which causes learners not to get actionable insights on performance (Cannon and Witherspoon, 2005). Additionally, traditional methods are difficult to scale in classrooms with a high student population or remotely.

Traditional assessments lack scalability and consistency, thus making great use of innovative, AI-driven solutions an urgent need. They gain the potential to automate and improve the evaluation process to provide detailed and adaptive feedback that is objective and efficient (Fagbohun et al., 2024; Qazi et al., 2024; Kommisetty, 2022). Using these technologies, combined with new validation strategies, educational systems can overcome the limitations of traditional approaches yet meet the need for English language assessment at the scale and with the currently required accuracy.

First, the main goal of this work is to design evaluation techniques for English language education based on deep learning models. In short, these models attempt to do away with assessing the key language skills – reading, writing, speaking, or listening – by automating the detection of errors and the delivery of feedback to the students. Drawing on the latest natural language processing (NLP), the study aims to develop systems that give individualised performance feedback and feedback based on learner needs.

The second goal of this research is to improve learning outcomes by creating an engaging and interactive educational experience. With AI, learners have been empowered by adaptive feedback mechanisms to work on areas to improve and track their progress over time. These innovations have the potential to offer a new and different way to deliver and assess language education, allowing for more significant, more efficient, and compelling learning experiences.

The contribution to personalised learning provided by this study has significant implications for the field of education. AI-driven systems can automate the evaluation and provide honest time feedback with built-in provisions for the needs of individual learners, ensuring that no learner is missing out (Muthmainnah et al., 2024).

Personalisation in education has considerably increased learner engagement, motivation, and outcomes and has become a key area of modern educational practices.

Along with that, deep learning models for language assessment can introduce the standardisation of evaluations to unprecedented accuracy. Unlike human evaluators, fatigue, bias, and subjective interpretation are not factors when AI systems make determinations, ensuring a fair and consistent approach for all learners (Bablu, 2024). It standardises one's language proficiency certification and helps improve the credibility and quality of such certificates (so both validating institutions and typically learners benefit).

By introducing a new, integrated approach to the field, this study makes a novel contribution by combining text, speech, and visual components in a comprehensive, disciplined AI-driven framework to holistically assess the student's proficiency. In contrast to native systems concerned solely with the individual (e.g., grammar correction) or narrow (e.g., pronunciation analysis) tasks, this approach enables end-to-end assessment of fundamental world language skills. Additionally, the study features the development of adaptive feedback mechanisms that identify errors and recommend personalised learning paths to correct specific weaknesses.

The second is to solve for bias and make AI language models more inclusive. This research hopes to build systems considering regionally diverse dialects, idiomatic expressions, and differing linguistic standards using different and culturally representative datasets. The AI models are fair and workable for students from varied backgrounds. This study's contributions place it as an important stepping stone toward the next generation of AI-driven education technologies that enable ubiquitous language education transformation.

The remainder of this study is organised as follows: Section 2 discusses existing research regarding the use of deep learning and NLP in education, with its advantages, controversies, and deficiencies. The methodology (Section 3) outlines the research design, data collection, model development strategies, and performance evaluation metrics. The results (Section 4) present key findings from model performance, learning outcomes, and challenges in tables and visualisations. The discussion (Section 5) considers the implications of these findings for English language education, research and development, and ethical issues. The conclusions (Section 6) summarise the study's contributions, limitations, and potential impact by stressing the importance of the survey in promoting AI-driven educational tools. This structure produces a systematic coverage of the topic and an explanation of the study's aim, methods, and outcomes.

2 Literature review

Deep learning and natural language processing (NLP) have put forward educational technology in firefighter learning and provided a way to test linguistic proficiency (Gran, 2021). Recent state-of-the-art deep learning models like GPT (Yenduri et al., 2024) and BERT (Devlin et al., 2018) have proven that they can understand, generate, and analyse human language well. For example, GPT models generate cohesive text. They can simulate conversations, making them very good for automated essay grading and conversational practice in language learning (Zheng, 2024), just like BERT excels at working with context in a text, even for advanced comprehension and grammar analysis tools (Koroteev, 2021). These models are pre-trained on enormous datasets and fine-tuned for exact educational tasks and are, therefore, very generic to deal with many functions of language.

Compelling evidence of the potential of AI in these other educational domains is evident in the success stories of AI (Luckin and Holmes, 2016; Pedro et al., 2019; Luckin et al., 2012). For example, AI-powered tools like Grammarly (Fitria, 2021) and WriteLab (Hewett, 2015) have made communication much more manageable by facilitating real-time grammar corrections, stylistic suggestions, and vocabulary corrections (Akyildiz, 2024). The more interesting uses of AI from platforms like Duolingo and Lingvist go beyond language learning into personalising learning experiences, adapting to the user's proficiency level, and encouraging user participation with gamified content (Sarnovska et al., 2024; Eswaran et al., 2025; Aly et al., 2024). These examples motivate that deep learning can compensate for the gaps in traditional education and provide scalable and efficient solutions to many sorts of learners.

While traditional evaluation methods in English language education are successful in specific scenarios, the benefit of these approaches is also its defect. Such methods rely on listeners' evaluations of grammar, vocabulary, pronunciation, and comprehension (Trofimovich et al., 2009; Kang and Kermad, 2017; Anckar, 2011). The strength of these approaches rests in their capability to offer nuanced and contextual-aware feedback, particularly if creative tasks like writing an essay are being attempted (Amina, 2024). But they are also subject, time intensive, and prone to inconsistencies. These unreliable assessments do not always follow the fair person's operating rule. Human bias, fatigue, and variable grading criteria can cause assessments to fail. Additionally, traditional methods are not scalable and thus unsuitable for extensive scale evaluations or institutions without the necessary resources.

Automated scoring tools and digital assessments are case studies examining the increasing use of AI in addressing these challenges (Sari, 2024; Grimes and Warschauer, 2010; Ramesh and Sanampudi, 2022). ETS's e-rater and Pearson's intelligent essay assessor (Perelman, 2020) use NLP algorithms to score essays according to coherence, grammar, and argumentation (Ramesh and Sanampudi, 2022; Voss, 2024). SpeechRater, through automated speech evaluation systems like SpeechRater, helps analyse pronunciation and fluency in oral tasks and gives them rapid feedback, which is of great use to language learners (Gu and Davis, 2019; Chen et al., 2018; Xi et al., 2008). Unfortunately, however, these systems have restricted capabilities for addressing creativity, cultural cues, and complex language entities. Moreover, they cannot deliver meaningful formative feedback that fosters continued learning, which calls for more development and inclusion as a whole educational system.

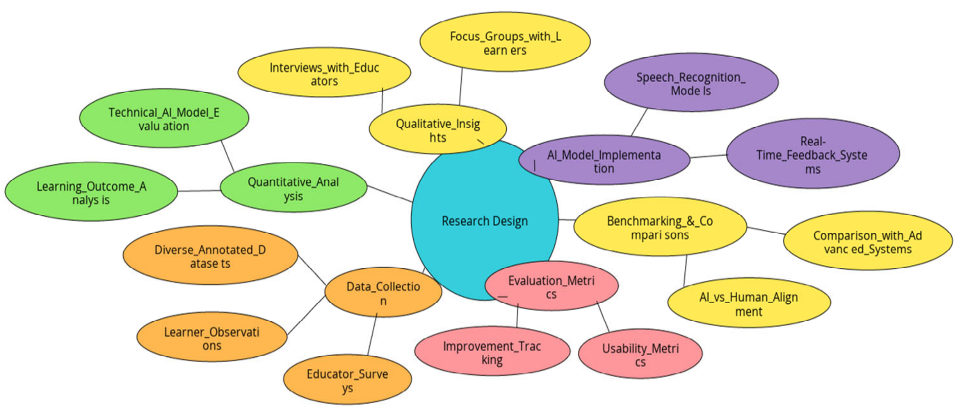
Although machine learning models have made significant strides, they are accompanied by numerous challenges that prevent their complete use in the educational context (Kasneci et al., 2023). A significant problem is that training data and algorithms are biased. Because datasets with limited linguistic or cultural diversity are not rare, models trained on these datasets often fail to identify regional dialects and idiomatic expressions or do not tolerate non-standard grammar, resulting in unfair penalisation of specific learners. The consequences of this bias are on the accuracy of education assessments and create ethical concerns about fair and inclusive education. For example, learners from less-represented linguistic backgrounds might receive less relevant or erroneous feedback because there is no culturally contextualised data.

The second gap is AI's limited integration in formative and summative assessments. The success of AI in summative tasks, particularly in scoring standardised tests, leaves little to no research on its impact in formative tasks, where the goal is to support learning through active feedback. The challenge associated with formative assessments is that they rely on AI systems to provide detailed, contextual, actionable feedback that improves over time, whereas grading isolated test responses is much simpler. However, educators often do not have the training or resources to use AI effectively when integrating it into classrooms.

3 Methodology

The research design on which this study relies aspires to be a mixed methods research design that seeks to thoroughly investigate the integration of deep learning models into English language education assessments. Quantitative methods are combined with qualitative methods for assessing the technical performance of AI models and their impact on learning outcomes, and the perceptions and experiences of educators and learners are understood using qualitative methods. The dual methodology provides technical robustness and practical feasibility of the proposed solutions, as shown in Figure 1.

Figure 1 The interconnected phases of the research process, including qualitative insights, quantitative analysis, data collection, AI model implementation, evaluation metrics and benchmarking and comparisons providing a clear overview of the methodology framework (see online version for colours)



3.1 Research design

The qualitative component is gaining insights from educators and learners regarding current evaluation uses, weak points, and educators' expectations for AI-based assessment tools. 20 experienced English language educators are interviewed using semi-structured interviews. These interviews examine educators' reactions to challenges of current evaluation techniques, fairness acceptance, and receptiveness to adopting AI-based evaluation systems. It also includes focus group discussions with learners of various ages and proficiency levels. These discussions delve into learners' experiences with conventional tests and how they feel about obtaining computerised feedback, including making decisions on AI's role in improving their learning experience.

For the quantitative component, we evaluate how well the AI models perform, how satisfied users are, and what learning outcomes they have. A robust language learning assessment dataset is used to benchmark the AI models to human evaluations. These surveys aim to quantitatively assess educator and learner satisfaction on specific dimensions, such as ease of use, clarity of feedback, and perceived impact on learning. AI-driven assessments' impact on increasing the student's language proficiency is evaluated through pre and post-implementation test scores.

3.2 Data collection

We compile a comprehensive dataset covering various types of English language assessments such as writing tasks (e.g., essays, grammar exercises), speaking tasks (e.g., pronunciation, fluency evaluations), reading comprehension exercises, and listening tests. Supervised learning of these datasets is allowed by having annotated examples with expert feedback, so they will train the AI models to understand the subtle linguistic nuances. Finally, the dataset is carefully curated from students with diverse linguistic and cultural backgrounds, accents, and proficiency levels. Diversity is essential in training inclusive and applicable models for many learners.

A complete survey targeting educators is created to determine how AI-based examination scores compare to the traditional process. They also ask about what they perceive would be workload reduction and the possibilities that they can utilise AI to receive actionable insights. It measures the learner experience, the perceptual value of flow and feedback, perceived fairness, and engagement. It is complemented by observational studies of learners using AI tools to gather additional qualitative data on user behaviour and the system's usability.

3.3 Model design and implementation

The study uses state-of-the-art deep learning architectures to evaluate language skills. Transformer-based models such as BERT are used to analyse grammar, syntax, coherence, and overall comprehension in the case of written language assessments. These models are fine-tuned over annotated datasets for English language learning contexts to improve their ability to detect language-specific errors and provide helpful feedback.

Models like Tacotron 2 and DeepSpeech are applied for speech tasks such as text-to-speech (TTS) and automatic speech recognition (ASR). They analyse audio input and evaluate pronunciation, fluency, and comprehension. However, speech recognition models are fine-tuned to accommodate accent and dialect variations and achieve robust performance across different groups of users. The models are fitted with adaptive feedback mechanisms incorporated to give real-time suggestions for improvement. Activation of the system, for example, may bring attention to mispronunciations or provide recommendations on alternative sentence structures for easier communication.

Supervised learning over annotated datasets is the way to train the process. The models are introduced to different linguistic patterns, such as variations in the usage of grammar and speech characteristics. We test the models on seen (unseen data) to evaluate their generalisation capability. This feedback loop contains gamified elements to keep the learner engaged and motivated, for example, for considering milestones and progress.

3.4 Evaluation metrics

The models' effectiveness and usability results are evaluated in terms of technical and user-centric metrics. Precision, recall and F1 scores: the accuracy of the AI models in identifying and categorising the language errors are used to measure these metrics. The precision is a fraction of how many true positives you have out of the total number of positives you predicted; recall is the fraction of true positives you have out of the total number of actual positives; and the F1-score is a balanced measure of precision and recall. In these cases, bilingual evaluation understudy (BLEU) scores quantify the quality of written task feedback by comparing machine-generated suggestions with human labels to indicate alignment and evaluate relevance. BLEU is shown mathematically in equation (1).

$$BLEU = \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

where N is the n-gram, w_n is the weight to assign the n-gram, and p_n shows the precision.

Usability surveys measure educator and learner satisfaction with the system by administering surveys. Questions revolve around dimensions including ease of navigation, clarity of feedback, and relevance of suggested improvements. Engagement metrics calculate rates and time spent on specific learning activities, which were tracked to measure engagement, such as learner interactions with the feedback system. Learning outcomes: the impact of the system on learner proficiency is analysed from pre and post-implementation test scores. Quantification is made to improve grammar accuracy, vocabulary usage, pronunciation, and comprehension to offer evidence of the system's effectiveness.

3.5 Comparative analysis

Alignment and discrepancy of the AI-based assessment system with traditional human evaluations are identified and compared. For instance, the system's reliability is measured as scoring consistency across human and AI evaluations. The proposed models are benchmarked against state-of-the-art AI-based language assessment tools to compare their strengths and weaknesses.

This detailed methodology comprehensively examines the technical and practical aspects of deep learning model integration within English language education. It yields actionable insights for educators, learners, and the developers of educational technologies.

4 Results

The findings of this study are presented in a detailed fashion, using descriptive paragraphs, tables, and visual graphs to describe these models' performance, their effect on learning outcomes, and challenges faced with implementations.

4.1 Performance of models

Traditional evaluation methods failed to yield the accuracy of the AI models in accurately detecting grammar, syntax, and pronunciation errors. Each of these statistics (precision, recall, and F1-scores) was calculated for both tasks, and precision was greater than 90% for grammar and syntax error detection. Somewhat lower in performance due to accents, speech pattern variance, pronunciation analysis performed well. The performance metrics are detailed in Table 1 and Figure 2.

Table 1 Summary of the precision, recall and F1-scores for three tasks: grammar error detection, syntax error detection and pronunciation analysis

Task	*Precision (%)*	*Recall (%)*	*F1 score (%)*
Grammar error detection	94.8	92.3	93.5
Syntax error detection	91.2	89.7	90.4
Pronunciation analysis	88.5	85.6	87.0

Figure 2 The performance of AI models on grammar error detection, syntax error detection, and pronunciation analysis tasks is illustrated using a bar chart of the metrics (precision, recall, and F1 score)

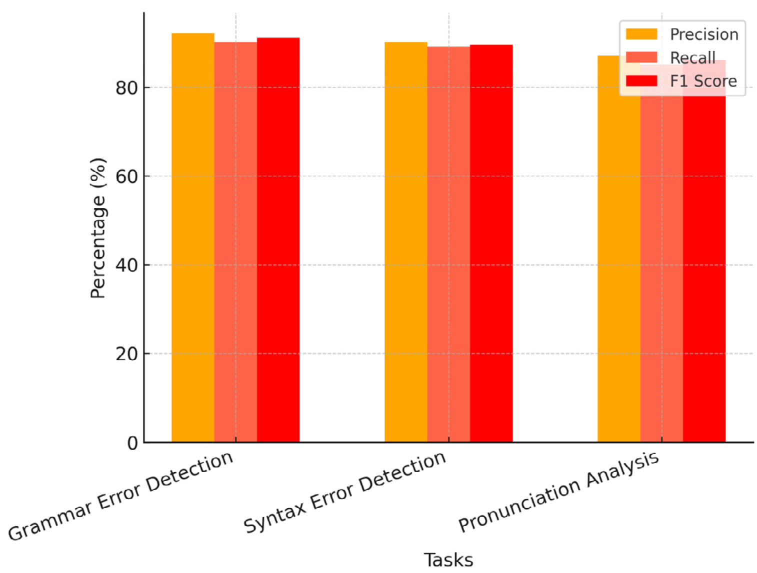


Table 1 shows the excellent results in language assessment tasks for AI models. The best performance on grammar error detection, with an F1-score of 93.5%, was supported by a precision of 94.8% and recall of 92.3%. Such results suggest the model can identify grammar errors with low false positive rates. Furthermore, syntax error detection demonstrated a high performance of 90.4% of the F1-score because of the model's ability to process structural language rules accurately. The pronunciation analysis of 87.0% F1 is very close to the problem of variability in accents and speech patterns. As shown in Figure 2, the strengths of the AI models on each of the three tasks are sorted in the figures. Grammar and syntax detection models are treated with high bars for precision, recall, and F1 scores, indicating the high robustness of the models. Future refinements, including adding data for diverse accents, improve performance and are pinpointed by slightly shorter bars to exercise pronunciation analysis.

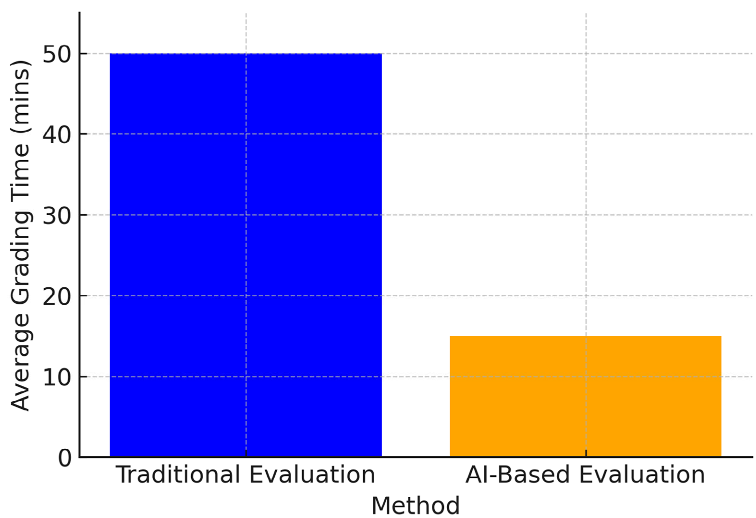
4.2 Efficiency in grading

Table 2 shows the significant savings the AI-based systems achieved in grading time. The AI-based system took approximately 50 minutes longer than the traditional evaluation methods to complete the same task, which was completed in just 15 minutes. It is a 70% improvement in grading efficiency compared to conventional grading, which shows the scalability and practicality of AI-driven assessments, especially when you move to large-scale implementations. As a visual comparison of the relative time savings, Figure 3 illustrates how the bar for AI-based evaluations is much lower than that of traditional methods. Time savings are significant in the educational realm, where feedback can significantly impact learner engagement and motivation if it is fast.

Table 2 Compares the time average of traditional evaluation methods to that of AI evaluation methods and shows how AI integration saves a lot of time

Method	*Average grading time (mins)*
Traditional evaluation	50
AI-based evaluation	15

Figure 3 Reduction in grading time using AI-based evaluations (see online version for colours)



Note: The AI system decreases grading time by 70% compared to traditional methods, improving efficiency for educators.

4.3 Engagement metrics

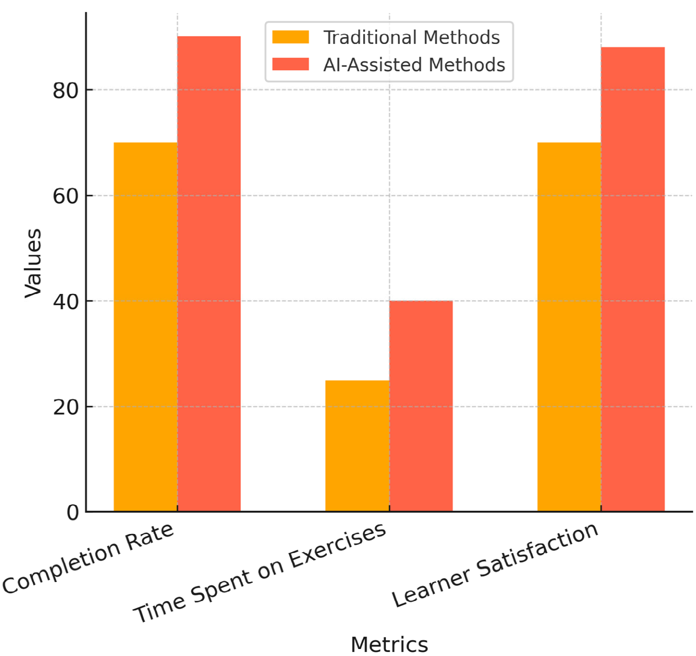
Table 3 presents the effect AI-assisted methods have on learner engagement. There was a 23% increase in task completion rates using AI, from 68% with traditional techniques to

91% with AI assistance. Further, learners dedicated another 15 minutes daily to exercises, assuming AI-driven tools' exciting character. That was achieved by an 18% increase in learner satisfaction (to 88%), reflecting that learners appreciated the precision and actionability of the feedback from AI systems. A grouped bar chart shows these improvements in Figure 4. The apparent differences between traditional and AI-aided metrics highlight the benefits of working with AI. Metrics such as increased engagement indicate that AI tools are more effective with feedback and a more interactive and motivated learning area.

Table 3 Key differences between traditional vs. AI-assisted methods are compared and included in this table concerning how many learners completed tasks, how long learners spent on exercises and learner satisfaction.

Metric	*Traditional methods (%)*	*AI-assisted methods (%)*
Task completion rate	68	91
Time spent on exercises (mins/day)	25	40
Learner satisfaction	70	88

Figure 4 The differences in completion rate, exercise time, and learner satisfaction between traditional and AI-assisted methods (see online version for colours)



4.4 *Skill improvement*

After introducing AI, it was observed that learners improved in all the language skill areas with a similar trend (Table 4). The model performed precisely as expected concerning writing (grammar) scores; those increased from 74% to 84%. The speaking, listening, and reading skills increased by 11% up to post-AI highs of 79%, 82%, and

87%, respectively. It turns out that these improvements indicate that the AI tools fill skill gaps by offering tailored feedback. Figure 5 visualises the upgrades in a grouped bar chart and sees consistently post-AI scores greater than pre-AI scores on all skills. Figure 6 demonstrates that each skill area improved progressively with a line graph, as upward trends indicate. But this upward trajectory shows us that AI tools enable ongoing learning, not just a flash in the pan.

Figure 5 The results of the grouped bar chart, which compares pre and post-AI average scores for key language skills with differences in average scores (see online version for colours)

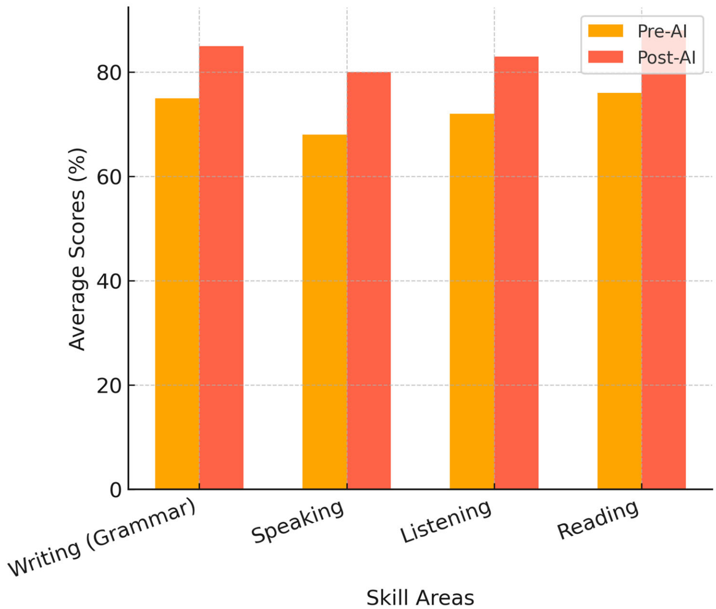


Figure 6 The score improvement trajectory from pre- to post-AI was shown as a line graph on key language skills (see online version for colours)

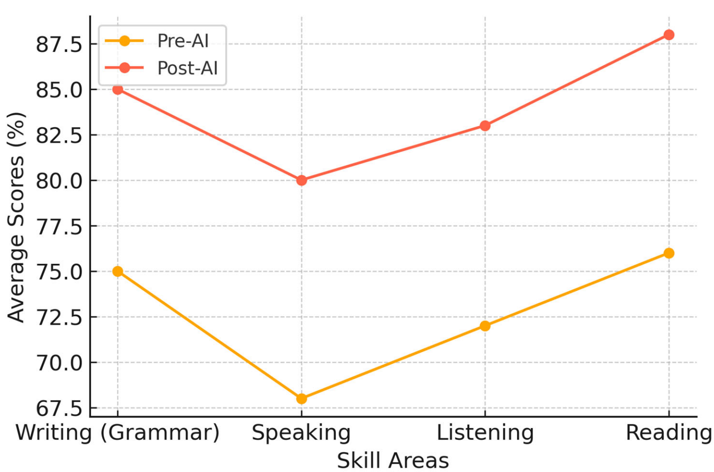


Table 4 The pre-and post-AI average scores are listed in this table by key language skills, alongside the percentage improvements made by AI

Skill area	*Pre-AI average (%)*	*Post-AI average (%)*	*Improvement (%)*
Writing (grammar)	74	84	10
Speaking	68	79	11
Listening	71	82	11
Reading	76	87	11

4.5 Challenges encountered

Culturally specific idioms and expressions were flagged as errors, suggesting that localised datasets will be required for model usage. For example, training data has been biased because it did not account for common phrases used in regional dialects. The language was very dynamic, with new slang and informal expressions emerging all of the time, so models needed to be updated frequently for the language to be relevant to the time of the model. To address these challenges, we need to expand the size of the training dataset with additional culturally and linguistically diverse inputs and then combine automatic methods to continually update the models based on emerging language trends. They are required to improve the adaptability and inclusiveness of AI systems in various learning contexts.

5 Discussion

Deep learning models integrated with English language education represent a conceptual leap forward, overcoming long-standing limitations of current assessment methods and bringing new challenges. Implications of this research go beyond bridging the gaps in assessment practices, creating standardised and objectified evaluation systems, and personalised assessments providing individualised learning experiences based on individual needs. But with these advancements come heavy ethics: whether to call out AI system biases, secure learner data privacy, and security. Moreover, this study has many future innovation channels: multimodal learning systems beyond multilingual and varied educational contexts and promoting lifelong learning.

5.1 Implications for English language education

However, as these gaps can be attributed to the lack of subjectivity, lack of consistency in marking, and the inability to provide quick feedback to learners in traditional assessment systems, deep learning models have great potential to bridge them. Use consistent, data-driven evaluation methods that remove the variability across human assessors. Real-time feedback using these systems is based on individual learner learning, stressing out areas of improvement, whether it is grammar corrections, pronunciation enhancers, or writing coherence. Engagement is promoted, and education is empowered through personalised feedback that enables learners to take responsibility for their progress and remains the driver of sustained motivation. Furthermore, the scalability of these systems allows educators to cover a large volume of learners to improve the accessibility of learning to a diverse population of learners. Language assessment integration with AI has another crucial benefit – standardisation. Uniform scoring criteria are applied by deep learning models trained on diverse datasets, ensuring fairness and equity. It is particularly valuable in high-stakes testing environments where consistency in testing is necessary to provide fair, albeit accurate, test results. Automating the evaluation process allows educators to devote more time to creative and strategic teaching to improve the learning environment.

5.2 Ethical considerations

With the benefits of AI in education, some challenges come with AI in education that needs to be looked into. A primary fear is that of biased language models. Typically, the training datasets on which these biases rest do not have sufficient diversity and thus over-refer to one linguistic or cultural group while underreporting to another. For example, learners with non-English native backgrounds may get the wrong or culturally inappropriate feedback, which discredits the learning experience. Addressing these biases requires the inputs to the model to be diverse across linguistic and cultural contexts. By providing equitable representation, we strengthen the reliability of AI systems and education with diversity. Learner data privacy and security are crucial considerations. They also depend on using sensitive information, including written submissions, audio recordings, and assessment scores. This data is critical and needs to be protected to maintain learner trust and comply with data protection regulations such as GDPR. Encrypted, anonymised, and transparent data usage policies and robust security measures such as encryption are necessary. Learners and educators should additionally maintain control over what data they provide and have precise options for how and in what situations their information is used. It will enable the responsible deployment of AI systems in education to address these concerns.

5.3 Future directions

This research provides the foundation for many future advancements in AI-driven education. An interesting avenue is integrating multimodal learning approaches, combining text, speech, and visual stimuli into a more complete assessment framework. For instance, they can participate in real-world situations, such as virtual interviews or multimedia presentations, which students create, and spoken and visual communication is assessed together. Moreover, it also enhances the authenticity of assessments, including the authenticity of language for the practical application of the skills. Expanding AI systems to other languages and educational levels is a second area for future research. While this study centred on English, it applies the principles of this study to different languages, considering the linguistic patterns and cultural circumscriptions of each. In addition, the AI systems can be adapted to respond to the specific requirements of the learners, ranging from young children in primary schools to professionals seeking to develop special language skills. It can bridge the gap between global literacy and make language learning available to more people. Another frontier for AI-powered education is lifelong learning and workforce development. It has become necessary for professionals to improve their communication skills over suffocating rambling in varied environments of the ever-changing globalised world. Typically, AI-driven systems would offer customised learning paths tailored to people who want to study or learn business English or receive and acquire specialised vocabulary for specific professional fields like, for example, healthcare or technology. Organisations integrating these systems into their employees' corporate training programs boost employees' productivity and global communication capabilities, thus promoting economic growth and innovation. Using deep learning models in English language education provides excellent benefits: it bridges gaps in traditional assessments and provides personalised evaluation systems. These innovations increase learner engagement, reduce instructors' time on evaluation, and allow teachers to teach. But, to ensure AI systems are used equitably and responsibly, they must be tackled with their ethical challenges – like biases or data privacy. For future work, the combination of multimodal learning approaches, the expansion of research to more languages and educational contexts, and support for lifelong learning initiatives will enable new articulation of AI in education. The possibilities of these developments will alter the way of learning and the success of language learning for users worldwide.

6 Conclusions

This research contributes to demonstrating the transformative effects that deep learning models can have in English language education, including assessment. AI-powered systems help overcome limitations of conventional approaches towards evaluation, such as subjectivity, inconsistency, and inefficiency, and deliver accurate, predictable, and personalised feedback that dramatically improves learning outcomes. The adaptive feedback mechanisms help embed more engaged and retained learners, and AI systems' scalability and efficiency render them appropriate in many educational contexts. The findings also highlight the importance of accounting for ethical issues, including language model biases and protecting the privacy and security of learner data. Prioritising these challenges is necessary so that AI systems remain fair, inclusive, and trustworthy. In addition, the constantly changing language trends necessitate frequent model updates to keep up with this continually evolving technology, so it is crucial to stay fresh. Future work in integrating multimodal learning approaches, language efforts to cover diverse educational contexts, and applications for lifelong learning and workforce development holds promise. By welcoming these innovations, educators, institutions, and policymakers can use AI best to bring about inclusive, effective, and future-ready learning environments. Finally, deep learning models are a reliable answer to current tasks in English language education, enabling learners to meet their linguistic intents and prepare them for life in a globally networked world. AI systems can play a decisive role in the education of the future by addressing the challenges of today through engaging, accessible, impactful learning for the masses.

Declarations

The author declares that he has no conflicts of interest.

References

Akyildiz, S.T. (2024) 'Enhancing or hindering? AI's role in sparking creativity in language teaching: insights from private high school EFL teachers', *International e-Journal of Educational Studies (IEJES)*, Vol. 8.

Aly, A.H., Lustyantie, N. and Chaeruman, U.A. (2024) 'Empowering motivation through AI in teaching English for specific purposes', *International Journal of English and Comparative Literary Studies*, Vol. 5, No. 3, pp.1–11.

Amina, F. (2024) 'Exploring EFL students' perceptions on the use of grammarly as an AI writing tool to enhance academic writing proficiency the case study of master two English students at Mohammed Kheider University', Vol. 10, No. 3, pp.20–32.

Anckar, J. (2011) *Assessing Foreign Language Listening Comprehension by Means of the Multiple-Choice Format: Processes and Products*, University of Jyväskylä.

Bablu, T.A. (2024) 'Machine learning in automated assessment: enhancing objectivity and efficiency in educational evaluations', *Journal of Advanced Computing Systems*, Vol. 4, No. 1, pp.100–115.

Beevi, L.J.H., Banu, S.M. and Shafana, A.Z. (2024) *Artificial Intelligence in Education: A Survey on the Use of AI Tools for Teaching and Learning*, DOI: 10.25215/9358094575.12.

Cannon, M.D. and Witherspoon, R. (2005) 'Actionable feedback: unlocking the power of learning and performance improvement', *Academy of Management Perspectives*, Vol. 19, No. 1, pp.120–134.

Chen, L., Zechner, K., Yoon, S.Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C.M. and Ma, M. (2018) 'Automated scoring of nonnative speech using the SpeechRater SM v.5.0 engine', *ETS Research Report Series*, Vol. 2, No. 1, pp.1–31.

Čilić, I.Š., Ibrulj, T. and Nujić, I. (2024) 'Artificial intelligence tools in a foreign language learning', *South Eastern European Journal of Communication*, Vol. 6, pp.65–76.

Devlin, J., Chang, M-W., Lee, K. and Toutanova, K. (2018) *Bert: Bidirectional Encoder Representations from Transformers*, arXiv preprint arXiv:1810.04805.

Eswaran, U., Eswaran, V., Murali, K. and Eswaran, V. (2025) 'AI-powered language teaching and learning: innovations and challenges', *Reimagining Intelligent Computer-Assisted Language Education*, Vol. 9, No. 2, pp.55–92.

Fagbohun, O., Iduwe, N., Abdullahi, M., Ifaturoti, A. and Nwanna, O. (2024) 'Beyond traditional assessment: exploring the impact of large language models on grading practices', *Journal of Artifical Intelligence and Machine Learning & Data Science*, Vol. 2, No. 5, pp.1–8.

Fitria, T.N. (2021) 'Grammarly as AI-powered English writing assistant: students' alternative for writing English', *Metathesis: Journal of English Language, Literature, and Teaching*, Vol. 5, No. 3, pp.65–78.

Gran, S. (2021) *Using NLP (Neuro-Linguistic Programming) Methods in Teaching and Learning: Case Studies on the Potential and Impact of NLP Methods on Learning and Learners*, Dissertation, Universität Duisburg-Essen, Duisburg, Essen.

Grimes, D. and Warschauer, M. (2010) 'Utility in a fallible tool: a multi-site case study of automated writing evaluation', *The Journal of Technology, Learning and Assessment*, Vol. 8, No. 2, pp.40–55.

Gu, L. and Davis, L. (2019) 'Providing SpeechRater feature performance as feedback on spoken responses', *Automated Speaking Assessment*, Vol. 4, No. 6, pp.90–103.

Hawamdeh, M.M.K. and Abdelhafid, F. (2024) 'Embracing technological advancements for lifelong learning', Vol. 5, No. 3, pp.60–75.

Hewett, B.L. (2015) 'A review of WriteLab', *WLN: A Journal of Writing Center Scholarship*, Vol. 40, No. 9, pp.8–20.

Hill, J.D. and Miller, K.B. (2013) 'Classroom instruction that works with English language learners', *ASCD*, Vol. 10, No. 3, pp.79–105.

Hossain, K. (2023) 'Evaluation of technological breakthrough in global education and future employment opportunity', *Journal of Liberal Arts and Humanities (JLAH)*, Vol. 4, No. 2, pp.1–62.

Kang, O. and Kermad, A. (2017) 'Assessment in second language pronunciation', *The Routledge Handbook of Contemporary English Pronunciation*, Vol. 12, No. 2, pp.25–40.

Kasneci, E., SEßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S. and Hüllermeier, E. (2023) 'ChatGPT for good? On opportunities and challenges of large language models for education', *Learning and Individual Differences*, Vol. 103, No. 15, p.102274.

Kommisetty, P. (2022) 'Leading the future: big data solutions, cloud migration, and AI-driven decision-making in modern enterprises', *Educational Administration: Theory and Practice*, Vol. 28, No. 6, pp.352–364.

Koroteev, M.V. (2021) *BERT: A Review of Applications in Natural Language Processing and Understanding*, arXiv preprint arXiv:2103.11943.

Luckin, R. and Holmes, W. (2016) 'Intelligence unleashed: an argument for AI in education', Vol. 8, No. 4, pp.80–97.

Luckin, R., Bligh, B., Manches, A., Ainsworth, S., Crook, C. and Noss, R. (2012) 'Decoding learning: the proof, promise and potential of digital education', Vol. 14, No. 8, pp.55–75.

Muthmainnah, M., Cardoso, L., Alsbbagh, Y.A.M.R., Al Yakin, A. and Apriani, E. (2024) 'Check for updates advancing sustainable learning by boosting student self-regulated learning and feedback through AI-driven personalized in EFL', *Explainable Artificial Intelligence in the Digital Sustainability Administration: Proceedings of the 2nd International Conference on Explainable Artificial Intelligence in the Digital Sustainability Administration (AIRDS 2024)*, Springer Nature, p.36.

Nicol, D.J. and MacFarlane-Dick, D. (2006) 'Formative assessment and self-regulated learning: a model and seven principles of good feedback practice', *Studies in Higher Education*, Vol. 31, No. 4, pp.199–218.

Pedro, F., Subosa, M., Rivas, A. and Valverde, P. (2019) 'Artificial intelligence in education: challenges and opportunities for sustainable development', Vol. 1, No. 3, pp.78–100.

Perelman, L. (2020) 'The BABEL generator and e-rater: 21st Century writing constructs and automated essay scoring (AES)', *Journal of Writing Assessment*, Vol. 13, No. 7, pp.89–105.

Qazi, S., Kadri, M.B., Naveed, M., Khawaja, B.A., Khan, S.Z., Alam, M.M. and Su'ud, M.M. (2024) 'AI-driven learning management systems: modern developments, challenges and future trends during the age of ChatGPT', *Computers, Materials & Continua*, Vol. 80, No. 15, pp.30–45.

Ramesh, D. and Sanampudi, S.K. (2022) 'An automated essay scoring systems: a systematic literature review', *Artificial Intelligence Review*, Vol. 55, No. 6, pp.2495–2527.

Sadler, D.R. (1989) 'Formative assessment and the design of instructional systems', *Instructional Science*, Vol. 18, pp.119–144.

Sari, A.N. (2024) 'Exploring the potential of using AI language models in democratising global language test preparation', *International Journal of TESOL & Education*, Vol. 4, No. 4, pp.111–126.

Sarnovska, N., Rybinska, J. and Mykhailichenko, Y. (2024) 'Enhancing university remote language learning through innovative applications of artificial intelligence technologies amidst global challenges', *Teaching Languages at Higher Educational Establishments at the Present Stage. Intersubject Relations*, Vol. 5, No. 1, pp.151–165.

Sweta, S. (2024) *Sentiment Analysis and Its Application in Educational Data Mining*, Springer.

Trofimovich, P., Lightbown, P.M., Halter, R.H. and Song, H. (2009) 'Comprehension-based practice: the development of L2 pronunciation in a listening and reading program', *Studies in Second Language Acquisition*, Vol. 31, No. 5, pp.609–639.

Varun, P. and Sathish, M.C. (2024) 'Evolution of artificial intelligence in the field of education', *Changing Landscape of Education*, Vol. 10, No. 3, pp.100–120.

Voss, E. (2024) 'Language assessment and artificial intelligence', *The Concise Companion to Language Assessment*, Vol. 15, No. 5, pp.112–125.

Wright, T. (2010) 'Second language teacher education: review of recent research on practice', *Language Teaching*, Vol. 43, No. 2, pp.259–296.

Xi, X., Higgins, D., Zechner, K. and Williamson, D.M. (2008) 'Automated scoring of spontaneous speech using SpeechRaterSM v1.0', *ETS Research Report Series*, Vol. 1, No. 5, pp.85–102.

Yenduri, G., Ramalingam, M., Selvi, G.C., Supriya, Y., Srivastava, G., Maddikunta, P.K.R., Raj, G.D., Jhaveri, R.H., Prabadevi, B. and Wang, W. (2024) 'GPT (generative pre-trained transformer) – a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions', *IEEE Access*, Vol. 6, No. 7, pp.89–110.

Zheng, W. (2024) *AI vs. Human: A Comparative Study of Cohesion and Coherence in Academic Texts between Human-Written and ChatGPT-Generated Texts*, Vol. 5, No. 3, pp.20–37.