JOURNAL OF COMMUNICATION, LANGUAGE AND CULTURE

Automated Evaluation of ESL Learners' English Writing Skills in English-Medium Instruction (EMI) through AI Writing Analytics

Waheed Shahzad[1,2*], Furrakh Abbas[3]

[1]Institute of Humanities and Arts, Khwaja Fareed UEIT, Rahim Yar Khan, Pakistan
[2]Department of English, Allama Iqbal Open University, Islamabad, Pakistan
[3]Centre for Languages and Translation Studies (CeLTS), Allama Iqbal Open University, Islamabad, Pakistan

*Corresponding author: waheed.shahzad@kfueit.edu.pk; ORCID iD: 0000-0002-2237-8237

ABSTRACT

This study analyses the use of artificial intelligence (AI)-driven automated writing evaluation (AWE) analytics and examines their role in enhancing the English writing skills of ESL learners in English-medium instruction (EMI) contexts. The review synthesises contemporary AWE systems, including those based on deep learning (DL), natural language processing (NLP), and generative AI approaches. It provides a detailed discussion of methods used for real-time feedback delivery, linguistic feature extraction, and the integration of AI-driven assessment with traditional teacher and peer feedback practices. In addition, the study critically evaluates empirical findings that highlight both the benefits and limitations of AI writing analytics in higher education EMI settings, including pedagogical, technical, and ethical challenges. Finally, the paper identifies future research directions and underscores the need for hybrid evaluation models that combine human oversight with automated systems to support technical accuracy, systematic assessment, and the development of higher-order writing skills in EMI contexts.

Keywords: AI writing analytics; automated writing evaluation; automated essay scoring; generative AI; English-medium instruction (EMI); ESL/EFL; feedback; formative assessment

Received: 16 June 2025**, Accepted:** 21 October 2025**, Published:** 29 January 2026

1.0 Introduction

The internationalisation of education and the rise of English-medium instruction (EMI) have brought specific expectations to bear on levels of English-language proficiency with respect to academic writing (Abbas & Bidin, 2022). In various EMI learning environments, students encounter the dilemma of gaining disciplinary knowledge and acquiring sufficient writing skills in English. Many conventional methods of evaluating writing that depend solely on teacher judgments can be expensive and prone to rater effects. As a response, automated writing evaluation (AWE) systems powered by AI writing analytics have emerged as a promising approach to provide timely and consistent individualised feedback for ESL learners (Tiandem-Adamou, 2024) with the aim of giving each learner what they need. This overview is focused on the current state of technology and research in automated EFL writing skills, focusing on such issues as AI integration for EMI feedback support, error correction, and revision history tracking.

Journal of Communication, Language and Culture
doi: https://doi.org/10.33093/jclc.2026.6.1.8

Published by MMU Press. URL: https://journals.mmupress.com/jclc

The remainder of this paper is organised as follows. Section 2.0 presents the background and related work on AI writing analytics and automated evaluation systems. Section 3.0 discusses selected methods and tools used in AI-based automated grading systems, including their architectures and feedback mechanisms. Section 4.0 consolidates the results of the empirical study and case analysis to identify the advantages and disadvantages of EMI environments. Section 5.0 describes key challenges and ethical issues and outlines potential directions for future research. Finally, Section 6.0 concludes the paper with remarks on the outlook of AI writing analytics in ESL writing development.

2.0 Literature Review

2.1 Automated Evaluation of ESL Learners' English Writing Skills

There has been increasing progress in AI-based writing tools over the past few years, which has been catalysed by developments in deep learning, neural network architectures, and NLP. Early AWE systems tended to concentrate on more surface-oriented characteristics of text, including grammar, syntax and vocabulary. However, modern systems have incorporated more linguistic analysis of documents to determine properties such as text coherence, structure and argumentation strength (Tiandem-Adamou, 2024; Wang, 2022). Examples include GenAI tools that have been developed for the promotion of ESL writing, offering immediate feedback on grammar corrections, vocabulary usage and overall writing coherence. Tiandem-Adamou (2024) study has also shown that the conjunction of generative AI with cooperative feedback strategies led to statistically significant improvements in the writing performance of EFL students learning in an EMI context.

The key novelty in these works is the integration of neural network models with non-artificially cut linguistic representations. Under early models, the focus was on manual feature engineering to extract surface level metrics, while recent deep learning models, including transformer-based architecture, have allowed for the extraction and review of dense semantic and syntactic features (Hussein et al., 2019). These tools are sometimes paired with feedback engines that produce actionable feedback on students' writing automatically. AI writing assistants are for more than catching typos. Research suggests that as students are given multiple opportunities to revise their work in accord with feedback, these tools can facilitate the learning of creating more reasoned, more coherent arguments and better organised essays (Chen, 2025).

Besides the studies in the context of English medium of instruction (EMI) which highlight AI as a valuable companion to contemporary teaching and learning approaches. As it is helpful to balance the inconsistencies in contemporary human grading. For instance, ChatGPT and Grammarly are effective tools to assess grammatical errors and deliver immediate responses by suggesting possible corrections. Both tools failed to achieve the great sides of writing, such as organisation, evaluating creativity and argument strength, and still need human feedback. The latter finding argues for a joint feedback model. This will allow teachers to focus less on the basic elements and more on providing a knowledgeable perspective: a hybrid model of sorts (Chen, 2025).

2.2 Conceptual Framework: AI-Based Writing Analytics in ESL/EMI Contexts

This section outlines three overlapping yet distinct technological paradigms—Automated Writing Evaluation (AWE), Automated Essay Scoring (AES), and Generative Artificial Intelligence (GenAI)—as they apply to English as a Second Language (ESL) learners in higher education English-medium instruction (EMI) contexts. Drawing on recent empirical and meta-analytic research, it differentiates the pedagogical functions and evaluative affordances of these technologies and argues for a hybrid human–AI feedback model that integrates automated support with instructor-mediated guidance.

2.2.1 AWE and AES: Definition and Evidence

Automated writing evaluations (AWE) systems aim at providing learners formative feedback by, for instance, tagging errors, suggesting improvements or helping in the revision of writing drafts. Whereas Automated Essay Scoring (AES) systems overlap with these, they are generally more widely structured around summative scoring strategies, involving holistic or analytic judgements assigned directly at the essay level using machine learning algorithms trained on human rated responses.

Findings from the meta-analysis indicate that AWE interventions in ESL/EFL contexts tend to produce a large effect size. For instance, Zhai and Ma (2022) reported a high effect size ($g = 0.861$, $p <. 001$) for writing AWE to improve across 26 primary studies ($N \approx 2{,}468$). In a three-level meta-analysis, Ngo et al. (2024) also found a moderate effect size ($g \approx 0.55$) of automated writing feedback on performance gains in writing. These findings indicate that AWE is more effective for post-secondary ESL/EFL students and for the genre of argumentative writing, but there are variations by intervention type and setting.

In EMI and ESL settings, AES can contribute to benchmarking and monitoring of proficiency but lacks the immediacy and detailed feedback necessary for drafting and revision. Thus, distinguishing AWE and AES are crucial for proper pedagogical design.

2.2.2 Generative AI (GenAI): Emerging Applications and Challenges

Generative AI (GenAI) tools, such as large-language-model systems that either generate, rewrite or scaffold text, offer new opportunities to expand writing support beyond error correction and scoring. Empirical (e.g., Mahapatra, 2024) shows that tools like ChatGPT can potentially be used as feedback tools or revision assistance in ESL tertiary contexts to support grammar, language structures, idea generation and writer's block. Some more exploratory research (Kim et al., 2025) has looked at students' views, noting the positive attitudes towards GenAI-assisted writing that were expressed but also raising issues of autonomy, authorship and academic integrity. In the area of writing analytics, Raitskaya and Tikhonova (2024) analysed 44 studies and found clusters related to generative text-production, scaffolded revision, and authorship/integrity.

Although GenAI has the potential to offer support for higher-order writing (e.g., rhetorical organisation, arguing, metacognitive prompts), the research is emerging. Students have also been observed to benefit in terms of active participation and modification when modifying GenAI-generated text, not if they just accept that.

2.2.3 Hybrid Human–AI Feedback Model for ESL/EMI Writing

Given the relative strengths and limitations of each paradigm, we therefore recommend a hybrid model in which automated writing evaluation (AWE) systems and generative AI tools support surface- and mid-level writing processes, while human instructors provide guidance on higher-order disciplinary, cultural, and rhetorical dimensions. Responsibilities are distributed across stages of the writing process as follows:

Planning: Gen AI supports brainstorming for outlines and ideas; the teacher specifies disciplinary genre, audience and criteria.

Drafting: AWE flags grammar, cohesion, vocabulary; GenAI returns rewriting suggestions and structure reminders; instructor checks the argumentative coherence, content relevance and genre adequacy.

Revising: AWE and GenAI are used to support multiple cycles of revision where peers and the teacher concentrate on clear content, disciplinary voice, critical thinking or metacognitive reflection.

Polishing & Summative Assessment: The AES can provide benchmarking; the instructor provides final feedback, checks it against institutional benchmarks and cultural-linguistic equity.

In this kind of hybrid ecology, AI language models become the "first line" of feedback for mechanical/linguistic issues and revision fluency. Whereas human feedback is still critical for disciplinary accuracy, metacognitive development, cultural/genre sensitivity and ethical oversight.

Such an integrated model responds to research signalling that automated feedback is most effective when combined with human mediation (e.g., effectiveness varies by context as shown in the meta-analyses) and that GenAI's benefits are maximised when learners engage actively (e.g., modifying AI outputs rather than accepting them without reflection).

2.3 Conceptual Distinctions between ESL and EFL Contexts

In the present work, ESL and EFL were maintained to reflect the jargon employed in the original studies. Although in applied linguistics these labels are sometimes used interchangeably, they reflect two distinct learning environments: ESL generally describes situations in which individuals work with a second language outside of school within the community, and EFL designates settings where English is primarily an academic subject and has limited usage in the larger society (Richards & Schmidt, 2013). Since the paper under consideration is inclusive of empirical studies carried out in different educational and sociolinguistic settings (e.g., English-medium universities in China and Turkey, tertiary institutions in Pakistan) by all terms accounting for fidelity to authors' original constructs approach a contextual specificity that parallels Kirkpatrick's framework (2008), it also maintains Sadeghpour and D'Angelo's models (Sadeghpour & D'Angelo, 2022). This inclusive language tracks onto the range of English learning environments that are found (and simply may or will be) within global EMI settings.

3. Methods

3.1 Research Design

This study is a literature review that synthesises prior research and does not generate new empirical data. It brings together the current state of knowledge on AI-supported writing tools and their use in assisting ESL students in English-medium instruction (EMI) classroom contexts. By synthesising, integrating, and critically evaluating findings from existing studies, the review offers a structured overview of contemporary research in this area.

Accordingly, this review provides several key contributions: (1) a timely overview of current evidence on automated writing evaluation in EMI settings; (2) identification of knowledge gaps and pedagogical challenges associated with the use of AI-supported writing tools for ESL learners; and (3) an analysis of the instructional affordances enabled by AI-based writing analytics within EMI environments. To achieve a comprehensive and balanced perspective, insights are drawn from multiple academic domains, including applied linguistics, education, and artificial intelligence, allowing for a more nuanced understanding of AI-mediated writing support in EMI contexts.

3.2 Sources of Data

The research that has been reviewed falls within the following sources and under the named broad categories: Journal articles in applied linguistics, computer-assisted language learning and AI in education, conference papers/proceedings to do with natural language processing (NLP) and educational technology, reports/policy documents about EMI (English as a medium of instruction) and English language learning, case studies on automated writing tools –e-rater, Grammarly; Write& Improve; AI-based feedback systems.

3.3 Selection Criteria

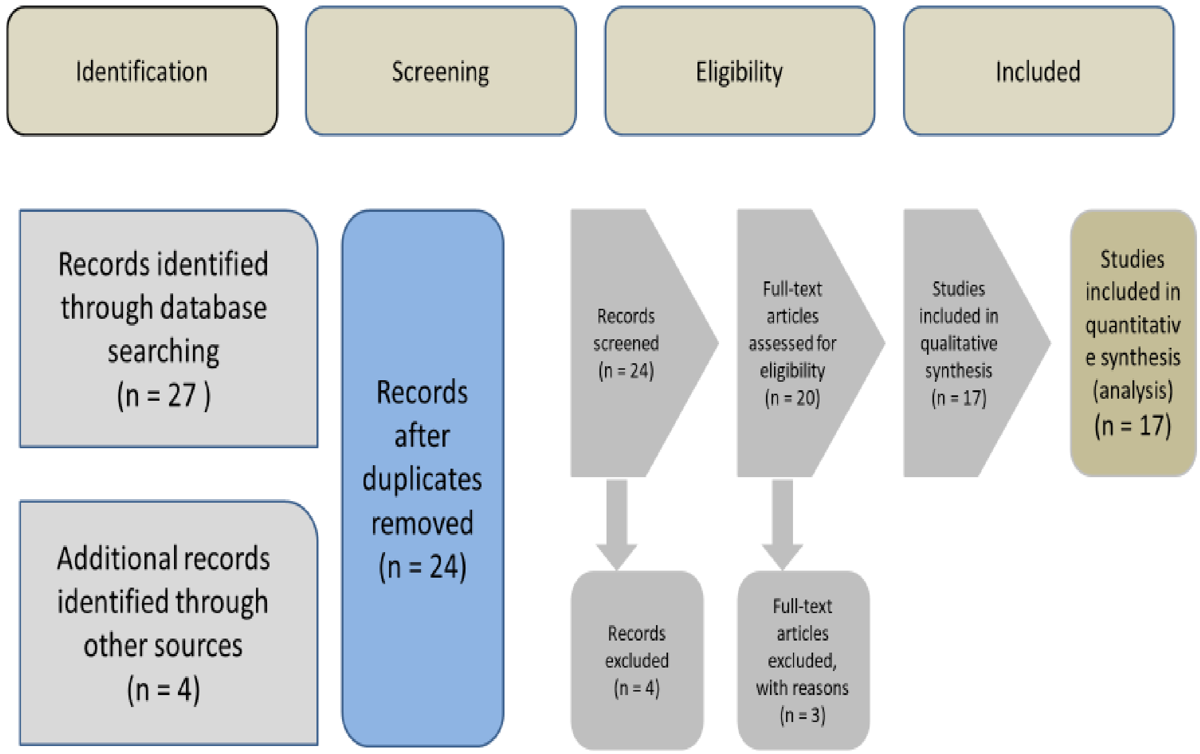
To ensure that the review was readable and that a reasonable focus was maintained, the following criteria were used (Figure 1). Figure 1 presents a conceptual PRISMA-style flow rather than a full systematic review diagram

Inclusion: Publication in English language studies between 2020 and 2025 that treat ESL/EFL writing, EMI, AI writing evaluation, or assessment.

Excluded: Publications in fields other than the language learning field, technical research of AI without an educational aspect, or published in other languages.

Figure 1

Selection of Publications for the review (PRISMA Flow Diagram)



3.4 Procedure

The review was a numerous step process: (A)Searching the online databases like Scopus, Web of Science, ERIC and Google Scholar using search terms including AI writing evaluation, ESL learners, EMI, writing feedback, and NLP in education. We also filtered the abstracts as well as the full texts for relevance related to the focus of this study. Categorising the literature into themes: (a) technology in writing assessment, (b) advantages and limitations for ESL learners, (c) AI role in EMI, and (d) ethical issues and fairness. Compared to analysing findings to understand trends, strengths, weaknesses and areas for future work.

The selection criteria for a research methodology utilizing bibliometric mapping (Figure 2) typically focus on three key dimensions highlighted by the visual clusters in the map:

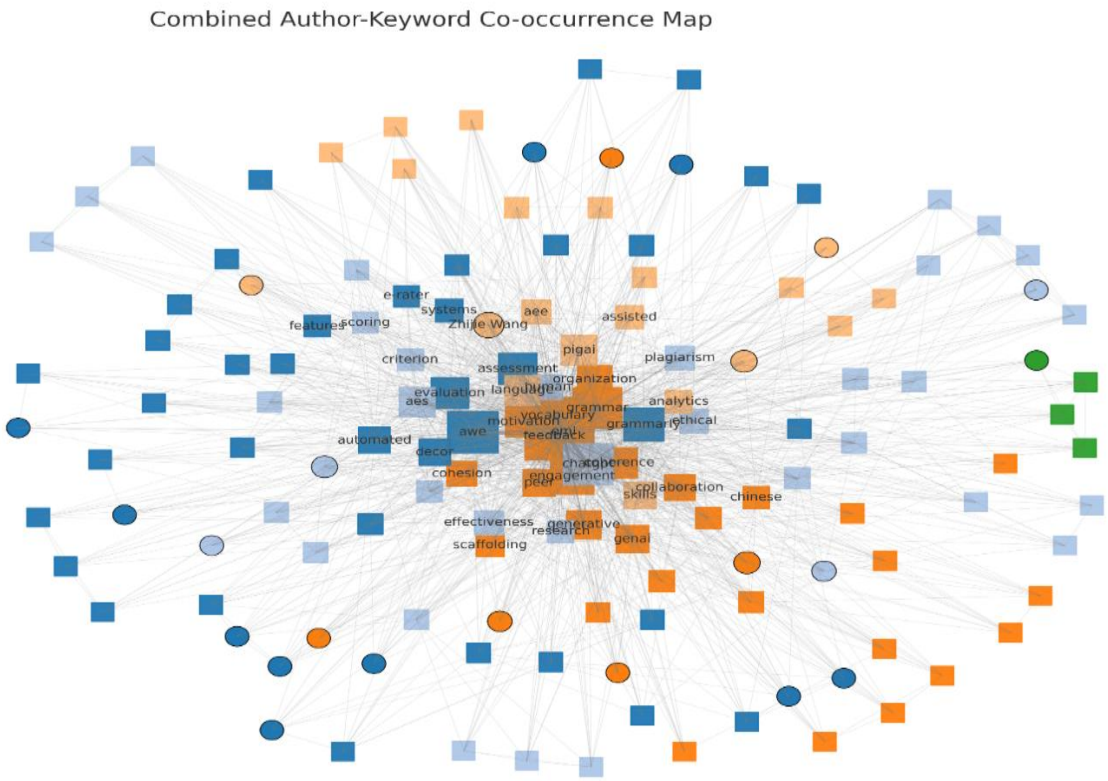
Keyword Relevancy: The central, dense orange cluster contains primary keywords such as "AWE" (Automated Writing Evaluation), "feedback," "grammar," "vocabulary," and "engagement." These terms act as the primary search strings and inclusion criteria. Only articles containing these keywords were selected to ensure the focus remained on the automated aspect of evaluation rather than general ESL pedagogy.

Technological Scope: The blue nodes dispersed around the center identify specific AI tools and analytical methods like "generative AI," "GenAI," "NLP" (Natural Language Processing), and "automated scoring systems" (e.g., e-rater, Criterion). The selection criteria likely used these terms to filter the literature for studies that specifically employ AI writing analytics the technical core of the article rather than traditional teacher-led feedback.

Contextual Filtering (EMI & ESL Scope): The map shows connections to "ESL learners," "EMI," and "Chinese" contexts (seen in the periphery orange and blue clusters). This indicates that the methodology's selection criteria were designed to be context-specific, prioritizing research that explores how AI tools perform specifically for ESL learners in EMI settings, where the linguistic demands of academic content add a layer of complexity to automated evaluation.

Figure 2

Co-occurrence Map



3.5 Data Analysis

Qualitative content review and analysis were used to analyse and interpret the selected literature. It focused on how AI writing analytics assisted in adopting/fostering the important aspects of ESL writing, such as grammar, vocabulary, fluency, coherence, and academic writing development across the EMI context. As the current study is a review, there is no ESL learner input. This constrains the extent to which the results can be generalised to common classroom use. Nevertheless, pooling a large number of studies in the review has allowed establishing evidence base for future research and practice in EMI settings (Appendix A).

Appendix A presents a structured overview of the eight research studies examined in this review on automated evaluation of ESL learners' writing skills within English-Medium Instruction (EMI) settings. The analysis compares these studies across key criteria, including research design, sample size, tools used, instruments, outcomes, and effect sizes. Together, they demonstrate how AI-driven writing analytics, automated writing evaluation (AWE) systems, and generative AI tools are being used internationally to assess and improve English writing performance among second-language learners.

a) Scope and Geographic Distribution: The reviewed studies span diverse contexts, including China, Pakistan, Turkey, and Iran. This geographic variation highlights the global rise of AI-supported writing evaluation in multilingual education systems. The inclusion of research from Pakistan provides a direct link to EMI contexts similar to the study's focus.

b) Research Designs and Participants: Sample sizes range from small qualitative datasets (e.g., 50 essays) to large quasi-experimental and comparative AWE evaluations involving more than 170 learners. Research designs include quasi-experimental studies, mixed-method approaches, corpus-based experiments, scoring validity studies, and exploratory process-tracing. This variation illustrates the methodological breadth used to validate AI tools in writing instruction.

c) AI Tools and Writing Analytics Platforms: A wide set of AI-powered tools appears across the studies, including generative AI (ChatGPT, GPT-4 mini), AWE systems (Pigai, iWrite, AWrite, Criterion) and hybrid feedback tools (Grammarly, DECOR system based on Detect–Explain–Rewrite). These tools support automated scoring, coherence evaluation, grammar checking, formative feedback, and revision tracking. Their use demonstrates how AI writing analytics can automate key components of ESL writing assessment, making them relevant to EMI environments where large class sizes often limit individualised feedback.

d) Instruments and Measured Outcomes: The instruments employed include writing pre- and post-tests, rubric-based scoring, motivation and engagement surveys, log data, teacher feedback records, and human vs AI scoring comparisons. Reported outcomes consistently focus on; writing proficiency and accuracy, coherence and organization, engagement, motivation, and revision behaviour, quality and reliability of AI-generated feedback. Together, these outcomes align with core competencies in EMI writing courses.

e) Key Findings and Effect Sizes: Across the studies, AI-supported writing analytics demonstrate significant positive effects on learner writing performance. Effect sizes range from moderate to strong, including correlations with human raters (e.g., $r = .81$, $r = .72$, $r = .67$) and improvements in writing scores (e.g., $t(98) = 3.45$, $p < .01$). These findings validate AI tools as reliable complements to human evaluation in educational settings.

f) Quality and Significance Notes: Most studies report strong empirical grounding, technical reliability, and alignment with contemporary CALL (Computer-Assisted Language Learning) frameworks. While some studies acknowledge limitations such as small samples or single-institution contexts, all report measurable improvements in writing quality, coherence, or learner engagement. This reinforces the potential for AI-driven analytics to enhance writing instruction under EMI conditions.

4. Results

4.1 AI-Based Writing Analytics in ESL/EMI Contexts

4.1.1 Natural Language Processing and Machine Learning Models

AI writing analytics tools are most often based on natural language processing methods and machine learning algorithms. Such systems, like the e-rater and Project Essay Grader, have traditionally been based on regression models that map a set of discrete linguistic features (e.g., vocabulary, grammar structures, and error frequencies) to holistic essay scores (Hussein et al., 2019). In contrast, modern approaches utilise neural architecture based on pre-trained language models, whose role is to produce context embedding of texts capturing local and global characteristics. For instance, more recent work using transformer models (including GPT-based model architectures) has shown state-of-the-art performance in capturing semantic nuances and discourse-level features that are crucial for assessing higher-order writing skills (Sajid et al., 2025).

The Detect, Explain, and Rewrite (DECOR) framework represents one such breakthrough, as it employs a multi-stage pipeline to detect, explain, and suggest rewrites for segments lacking coherence in learner texts (Sajid et al., 2025). Using large corpora such as EF-Cambridge Open Language Database (EFCAMDAT), it is possible to obtain a robust assessment and feedback that aligns with human judgment. Furthermore, by integrating self-attention mechanics within these models, the capacity to assess parts of the essay texts is enhanced, allowing for a holistic evaluation of text in terms of grammar, logic, and writing style (Sajid et al., 2025; Wang, 2022).

4.1.2 Automated Feedback Generation and Personalised Learning

Today's AI writing analytics are created to give immediate feedback, and they're personalised. The systems, which deliver real-time processing of learners' submissions, return corrections and suggestions to be used to iteratively and interactively revise the text. AI feedback in these tools typically addresses three main areas: (i) mechanical elements, including grammar and lexico-grammar repairing; (ii) lexical enrichment through lexis suggestions given by (disambiguated) central vocabulary, and (iii) structural improvements enabling better coherence and logical development in writing (Chen, 2025; Rahman et al., 2023).

Personalised feedback is achieved by adapting responses to the learner's individual errors and proficiency level. In EMI settings where class sizes can be large and individual teacher attention is limited, AI writing analytics offer scalable solutions that provide consistent and detailed responses across a diverse student body (Rahman et al., 2023). Systems such as those developed by Tiandem-Adamou (2024) have successfully demonstrated that providing immediate AI-generated feedback can significantly improve ESL learners' overall writing performance while reducing the cognitive load associated with manual error correction (Tiandem-Adamou, 2024).

4.1.3 Integration with Cooperative and Instructor-Led Feedback

Despite the technical sophistication of AI writing analytics, several studies emphasise the importance of integrating automated feedback with human evaluation. Research shows that AI tools are great for catching obvious mistakes and fixing grammar, but they often miss the bigger picture. They struggle to understand context or cultural nuance, which makes it hard for them to give useful feedback on complex elements like a writer's argument or unique narrative voice (Chen, 2025). As a result, many EMI rooms are now embracing blended learning. The approach combines AI-generated corrections and manual feedback, in traditional form from teachers or peers.

The advantage of this three-pronged approach is that it leverages the strengths of each. To mechanical errors, the AI contributes with its speed and consistency; to the deeper pedagogy insight of human interaction, teachers still contribute (Chen, 2025). For instance, it has been reported that students are most satisfied when the automatic response by AI is combined with a team writing workshop or a teacher's individual feedback on their outputs (Song & Song, 2023). This joint coherence enables both sides to learn.

4.2 Empirical Findings and Case Studies

4.2.1 How AI Feedback Improves Writing Skills

Increasing evidence suggests that AI writing tools lead to actual and measurable learning gains among ESL students enrolled in EMI programs. In a key 2024 study, Tiandem-Adamou conducted a controlled experiment for students using AI writing assistants and others because of traditional methods only. The AI-assisted group showed significantly greater improvements in areas such as organisation of ideas, precision of expression and overall writing clarity. Their mean performance improved dramatically, in sharp contrast to the slight overall improvement found for the control group (Tiandem-Adamou, 2024). This suggests that immediate, specific feedback from AI is very effective in showing students how to revise and grow academically.

Other studies also support these findings, which relate AI tools to better linguistic correctness and more difficult sentence building. For instance, the 2022 study of Zhijie Wang, for example, demonstrated that AI evaluation systems accurately evaluated such linguistic features as vocabulary use, sentence variability and technical mechanics in addition to increasing students' motivation and self-esteem (Wang, 2022). The other advantage is the system's capacity to rapidly and consistently score many essays, thereby ensuring a degree of uniformity in grading that most human assessors find difficult to achieve because their subjective standards are often uneven (Hussein et al., 2019; Pratama & Sulistiyo, 2024).

4.2.2 Automated Scoring: How AI and Human Grading Compare

Artificial Intelligence (AI) based models, such as Automated Essay Scoring (AES), are examining essays, where their performance is commonly judged by a concordance with human teacher grading. This mixed picture¬ is also reported in literature concerned with the tools similar to ChatGPT-based graders: while AI graders agree in general with human perceptions, they do not always agree all the time. The largest discrepancies occur in assessments of higher-order features of writing, including rhetorical quality, argumentative strength, and creative expression (Bannister et al., 2023; Uyar & Büyükahıska, 2025).

This disconnection is evident in a study by Uyar and Büyükahıska (2025), where the language features of B2-level learner essays were examined. They found that feedback from the AI, even one with a final score close to that of a teacher, was often generic and difficult to understand, or unrelated to the training example. This led students to display the comment to a teacher (if available) and ask them which in turn showed that the AI reasoning process is less nuanced than a human grader's explanation (Bannister et al., 2023) Ultimately, these studies show that while AES systems excel at efficiently assessing grammar and mechanics, the accurate evaluation of content and ideas still requires human judgment.

4.2.3 Enhancing Learner Engagement and Motivation

The benefits of AI writing tools are not just about better grammar; they also play a significant role in boosting student motivation. Research suggests that because AI provides feedback instantly, it helps ease the anxiety students often feel waiting for a teacher's evaluation. This can lead to a more positive outlook on revising their work and learning in general (Chen, 2025; Song & Song, 2023). Students also report valuing the detailed, personalised suggestions they receive, which allow them to progressively improve through multiple drafts. This iterative process helps learners feel a greater sense of ownership and control over their own progress (Song & Song, 2023; Tiandem-Adamou, 2024). Furthermore, the integration of AI tools into cooperative learning environments, where students work collaboratively and share feedback insights, has been shown to enhance peer interactions and overall classroom participation (Tiandem-Adamou, 2024).

5.0 Discussion

5.1 Benefits of AI Writing Analytics in EMI Contexts

The benefits of using an AI writing tool in EMI are many and dynamic. For starters, these tools are an incredible time saver for teachers. By automatically flagging frequent errors and providing students with immediate feedback, they can also cut down on the time required to grade. This allows teachers in large classes, where it's not always easy to offer individualised help, to reclaim hours of their day (Rahman et al., 2023; Tiandem-Adamou, 2024). One other major benefit is that they are truly objective. Contrastingly to humans, who can be influenced by unconscious bias, AI systems provide consistent, unbiased international standards for all students and therefore support a safer assessment (Hussein et al., 2019; Tiandem-Adamou, 2024). "Institutionally, AI also lends itself to great scalability." They can be easily scaled down to any number of classrooms at very little additional cost and are therefore an excellent choice for schools and universities that need to maintain a balance between quality and economy (Tiandem-Adamou, 2024). Perhaps most importantly, today's AI tools can customise the learning experience. They adjust to the needs of individual students, allowing learners to work at their own pace and concentrate on areas where they are weakest. This facilitates SRL and is particularly beneficial in multi-lingual EMI classes where learners are at different language levels (Chen, 2025; Song & Song, 2023).

5.2 Challenges and Limitations

Despite their clear benefits, AI writing tools still face significant challenges. A major limitation is their struggle to evaluate the core of good writing: strong arguments, clear organisation, and creative expression. While they are excellent at catching grammatical mistakes, they often miss the nuance of how ideas are developed and connected, making human oversight essential (Chen, 2025).

This leads to broader ethical concerns. Some educators worry that over-reliance on these tools might stifle a student's ability to think critically and solve problems independently (Rahman et al., 2023). Others raise serious questions about fairness, pointing out that AI systems can be culturally insensitive or perpetuate biases hidden in their training data, which could disadvantage some student groups (Rahman et al., 2023; Van Wyk, 2025).

The literature also underscores significant limitations in evaluating higher-order writing skills. While AI tools have become quite skilled at catching grammar mistakes and suggesting better vocabulary, they still face a major hurdle: evaluating the heart of good writing. Tasks that require judging a paper's ideas, how they are organised, or its creative flair remain a significant challenge for many systems. Even with advances in machine learning, AI often misses the subtleties needed to assess these complex, high-level skills (Chen, 2025).

This matches what many students report. They find AI feedback incredibly useful for fixing surface-level errors and polishing their word choice but often find it unhelpful for improving the overall flow of their essay or the strength of their central argument (Chen, 2025). Their essay or the power of their master argument (Chen, 2025). Therein lies the danger, where if everyone puts all their faith in AI alone, writers may start to lose sight of the key factor that sets powerful writing apart so vastly, which are critical thinking and originality. The answer, then, is not to throw out AI but to combine it with human expertise. This balanced approach guarantees that students get help on all sentence level content (Chen, 2025).

On the technical side, reliability remains a hurdle. Studies show that AI scoring can be inconsistent and does not always align with human grades, especially across different essay genres or proficiency levels (Bannister et al., 2023; Uyar & Büyükahıska, 2025). Put another way, AI is not a simple plug-and-play application. The problem is that it requires precise tuning and constant testing, in order to make sure these tools are fair, accurate, and moving in the right direction with respect to education (Wang, 2022).

5.3 Future Directions and Research Needs

Second, longitudinal studies are needed to investigate the long-term effect of AI-feedback on ESL. Learners' writing in EMI contexts and looking at whether learners are motivated to write in an additional language. Although many studies have shown promising short-term effects, it is unclear whether such improvements are maintained over time and how they affect learners' do more of their own writing (Song & Song, 2023). Going forward, a major area for research is determining the most effective ways to combine AI tools with input from teachers and peers. The best systems will probably be those that are hybrid and strike a careful balance, using technology to handle some tasks while relying on human insight for others. It has the potential to mitigate vulnerabilities of humans and machines, while leveraging their respective strengths (Chen, 2025).

Yet another is to make AI more aware of culture. For such tools to be effective across the broad spectrum of EMI classrooms, they must respect all students' linguistic and cultural roots. Doing so will require AI to be trained on far more diverse datasets and be programmed to identify locally relevant language patterns. Such solutions will allow it to return feedback not just accurate, but suitable and useful in various individual learning settings (Charpentier-Jiménez, 2024; Chen, 2025). Lastly, we cannot forget about moral issues. And as they do, teachers and developers will have to engage on such issues as data privacy and the potential for cheating, along with avoiding having students become over-reliant on automated help. Clear rules and guidelines will need to be set in place to make sure these powerful tools help, not hinder, teaching (Chen, 2025).

5.3.1 Hybrid Human–AI Feedback Model

Based on the analysis and discussion, the hybrid human-AI feedback model is developed. The visual depiction of roles across stages. The concise explanation and a sample rubric showing AI vs Human responsibilities.

Table 2 presents a hybrid Human AI feedback model that outlines the distribution of tasks between AI-driven writing analytics (such as AWE systems and generative AI tools) and human evaluators (teachers or peers). The table categorises key writing aspects and clarifies which components can be reliably automated and which require human judgment. This model reflects current evidence from the literature and supports practical decision-making for English-Medium Instruction (EMI) writing courses.

a) AI-Suitable Tasks: The table shows that AI tools are well-suited for tasks involving pattern recognition, surface-level corrections, and automated text analysis. These include; grammar and mechanics such as punctuation, typo detection, and basic syntactic corrections, vocabulary assistance through synonym suggestions and collocation checks, coherence and organisation via outline generation and paragraph-level cohesion guidance, content support through factual surface checks and suggested citation wording, genre awareness by generating sample structures in relevant academic genres, and metacognitive prompts that encourage reflection and provide revision strategies. These tasks rely on AI's ability to process large datasets, detect repeated patterns, and generate structured suggestions quickly. In EMI contexts where class sizes are large, these automated functions help reduce teachers' workload and give learners immediate, formative feedback.

Table 2

Hybrid Human–AI Feedback Model: Tasks AI can handle vs tasks requiring human feedback

Writing Aspect	AI (AWE/GenAI) - Suitable Tasks	Human (Teacher/Peer) - Essential Tasks	Example Criteria / Notes
Grammar & Mechanics	Error detection; suggested rewrites; punctuation	Nuanced syntactic appropriateness; register; teaching moments	AI flag % errors; teacher checks contextual use
Vocabulary & Lexical Choice	Suggest synonyms; collocation checks	Discipline-specific term choice; nuance, connotation	AI suggestions as options; teacher approves domain terms
Coherence & Organization	Outline generation; paragraph-level cohesion cues	Argument structure; thesis development; evidence quality	AI provides suggested reorganisations; teacher evaluates rhetorical moves
Content Accuracy & Evidence	Can surface-check facts; suggest citation wording	Verify claims; assess evidence quality and interpretation	Humans must validate sources and disciplinary claims
Audience & Genre Awareness	Generate examples in the target genre, mimic style	Ensure genre conventions and cultural sensitivity	The teacher provides genre-specific rubrics
Metacognitive Prompts	Generate reflection questions; revision suggestions	Guide learners through reflective responses; mentoring	Combine AI prompts with teacher scaffolded activities

b) Human-Essential Tasks: The table also clarifies the limits of AI, identifying areas where human insight is essential. These include; nuanced syntactic appropriateness, register control, and context-specific language choices, discipline-specific vocabulary that requires expert validation, argument development and critical interpretation, such as assessing thesis quality, cohesion across sections, and the strength of evidence, accuracy and validity of claims, particularly when evaluating sources or disciplinary content, audience and cultural sensitivity, which AI cannot reliably assess, and mentoring and reflective guidance, which depend on pedagogical understanding. These areas require contextual awareness, cultural knowledge, and academic judgement capabilities that AI tools cannot fully replicate. Teachers and peers provide the evaluative depth and disciplinary expertise needed to support higher-order writing skills.

c) Integration Notes: The final column offers practical criteria for combining AI and human feedback. Examples include using AI to flag grammatical errors while teachers check contextual meaning, allowing AI to suggest synonyms but relying on teachers to validate discipline-specific terms, using AI-generated outlines while teachers refine argumentation, and combining AI-generated prompts with teacher-led reflective activities. These notes emphasise that AI feedback serves as a starting point, while teachers shape, validate, and personalise the learning process. This creates a balanced approach where technology enhances efficiency without replacing professional judgement.

6.0 Conclusion

In summary, AI evaluation systems are by no means an ideal substitute for a teacher's judgment, but they do appear to offer a practical, efficient and scalable mechanism for assisting with the teaching of writing skills to ESL students. There will be more work to do for all of us in education, on all these tools, but a constant learning loop and doing more R&D on best practices ultimately should pay dividends.

6.1 Recommendations

Instructors should use AWE tools as an initial source of formative feedback while keeping summative evaluation in the hands of human ratters. This balance allows students to benefit from quick, surface-level guidance while ensuring that final judgement reflects nuanced academic expectations. To make this process effective, instructors also need to provide explicit AI-literacy support and guide students through rubric-based revision activities, so they understand how to interpret and act on automated feedback. At the programme level, institutions should develop clear policies that outline appropriate use of AI, address data privacy concerns, and ensure that assessment practices remain aligned with curriculum goals. Ed-tech developers have a parallel responsibility to design feedback systems that are transparent, easy to interpret, and grounded in principles of explainability, while also offering options for exporting or anonymising student data. For researchers, the main recommendation is to prioritise longitudinal and cross-cultural studies in EMI contexts so that stronger evidence can be generated on the effectiveness and limits of AI-supported writing evaluation.

6.2 Future Research Directions

Future research should focus on long-term randomised controlled trials in EMI higher education to determine whether improvements supported by AI tools remain stable over time. There is also a need to develop culturally responsive feedback corpora and to examine how bias may appear in AI comments provided to writers who use English as an additional language. Another important direction is to evaluate the validity of AI feedback across different genres and proficiency levels in order to identify the contexts in which automated feedback is dependable and the areas where human judgement continues to be essential.

Conflict of Interest

The author(s) have declared that no competing interests exist.

Author Contribution Statement

WS: Conceptualization, Data Curation, Methodology, Validation, Writing – Original Draft Preparation.
FA: Project Administration, Supervision, Writing – Review & Editing.

Funding

This research received no external funding.

Ethics Statement

This research did not require IRB approval because it did not involve human participants.

Data Access Statement:

Research data supporting this publication are available upon request from the corresponding author.

Author Biography

Waheed Shahzad is a Lecturer in Institute of Humanities and Arts, Khwaja Fareed University of Engineering and Information Technology Rahim Yar Khan. He is also a Ph.D. Scholar of English Linguistics at Allama Iqbal Open University Islamabad. His major research work is in the field of Language Planning and Policy.

Furrakh Abbas is currently serving as Assistant Professor in Centre for Languages and Translation Studies (CeLTS), Allama Iqbal Open University (AIOU) Islamabad. With two PhDs to his credit, first PhD in Linguistics from UMT, Pakistan and second PhD in Applied Linguistics from UUM Malaysia and more than 15 years of experience in academia, training and research, he specializes in fields of language planning and policy, applied Sociolinguistics and integration of GenAI in language documentation and preservation.

References

Abbas, F., & Bidin, S. J. (2022). A critical analysis of the language planning and policy (LPP) in Pakistan and its impact on indigenous languages of Pakistan. *Eurasian Journal of Applied Linguistics, 8*(1), 85–96. https://ejal.info/menuscript/index.php/ejal/article/view/135

Bannister, P., Santamaría-Urbieta, A., & Alcalde-Peñalver, E. (2023). A Delphi study on generative artificial intelligence and English-medium instruction assessment: Implications for social justice. *Iranian Journal of Language Teaching Research, 11*(3), 53–80. https://doi.org/10.30466/ijltr.2023.121406

Charpentier-Jiménez, W. (2024). Assessing artificial intelligence and professors' calibration in English as a foreign language writing courses at a Costa Rican public university. *Actualidades Investigativas en Educación, 24*(1), 533–557. https://doi.org/10.15517/aie.v24i1.55612

Chen, Q. (2025). Students' perceptions of AI-powered feedback in English writing: Benefits and challenges in higher education. *International Journal of Changes in Education*. https://doi.org/10.47852/bonviewIJCE52025580

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, e208. https://doi.org/10.7717/peerj-cs.208

Kim, J., Yu, S., Detrick, R., & Li, N. (2025). Exploring students' perspectives on generative AI-assisted academic writing. *Education and Information Technologies, 30*(1), 1265–1300. https://doi.org/10.1007/s10639-024-12878-7

Kirkpatrick, A. (2008). *World Englishes: Implications for international communication and English language teaching*. Cambridge University Press.

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed-methods intervention study. *Smart Learning Environments, 11*(1), 9. https://doi.org/10.1186/s40561-024-00295-9

Ngo, T. T. N., Chen, H. H. J., & Lai, K. K. W. (2024). The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. *Interactive Learning Environments, 32*(2), 727–744. https://doi.org/10.1080/10494820.2022.2096642

Pratama, A., & Sulistiyo, U. (2024). A systematic review of artificial intelligence in enhancing English foreign learners' writing skills. *PPSDP International Journal of Education, 3*(2), 170–181. https://doi.org/10.59175/pijed.v3i2.299

Rahman, N. A. A., Zulkornain, L. H., Mat, A. C., & Kustati, M. (2023). Assessing writing abilities using AI-powered writing evaluations. *Journal of Asian Behavioural Studies, 8*(24), 1–17. https://doi.org/10.21834/jabs.v8i24.420

Raitskaya, L., & Tikhonova, E. (2024). Applications of generative AI-powered language tools in academic writing: A scoping review. *Journal of Language and Education, 10*(4), 5–30. https://doi.org/10.17323/jle.2024.24181

Richards, J. C., & Schmidt, R. W. (2013). *Longman dictionary of language teaching and applied linguistics*. Routledge.

Sadeghpour, M., & D'Angelo, J. (2022). World Englishes and "Global Englishes": Competing or complementary paradigms? *Asian Englishes, 24*(2), 211–221. https://doi.org/10.1080/13488678.2022.2076368

Sajid, A., Amjad, B., & Khan, S. (2025). Enhancing second language writing assessment through natural language processing: A corpus-based study. *Journal of Applied Linguistics and TESOL, 8*(3), 671–682. https://doi.org/10.63878/jalt1019

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1260843

Tiandem-Adamou, Y. (2024). Using generative artificial intelligence to support EFL students' writing proficiency in university in China. *Journal of Educational Technology and Innovation, 6*(4). https://doi.org/10.61414/jeti.v6i4.213

Uyar, A. C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. *International Journal of Assessment Tools in Education, 12*(1), 20–32. https://doi.org/10.21449/ijate.1517994

Van Wyk, M. M. (2025). Postgraduate students' voices on leveraging Grammarly as an AI-powered tool in academic writing. *Journal of Education (University of KwaZulu-Natal)*, (98), 103–123. https://doi.org/10.17159/2520-9868/i98a06

Wang, Z. (2022). Computer-assisted EFL writing and evaluations based on artificial intelligence: A case from a college reading and writing course. *Library Hi Tech, 40*(1), 80–97. https://doi.org/10.1108/LHT-05-2020-0113

Yang, H., Gao, C., & Shen, H. Z. (2024). Learner interaction with, and response to, AI-programmed automated writing evaluation feedback in EFL writing: An exploratory study. *Education and Information Technologies, 29*(4), 3837–3858. https://doi.org/10.1007/s10639-023-11991-3

Zhai, N., & Ma, X. (2022). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research, 61*(4), 875–900. https://doi.org/10.1177/07356331221127300

Appendix A

Table 1

Summary of the Analysis

Criteria	**Detailed Information in Appendix B**							
Author	Tiandem-Adamou, Y.	Wang, Z.	Song, C., & Song, Y.	Chen, Q.	Uyar and Büyükahıska	Sajid, M., Amjad, R., & Khan, S.	Yang, L., Gao, Y., & Shen, M.	Link, S., Mehrzad, M., & Rahimi, M.
Year	2024	2022	2023	2025	2025	2025	2024	2020
Country	China	China	China	China	Turkey	Pakistan	China	Iran
N	100	178	50	Not reported	50 essays	Corpus-based (EFCAMDAT)	Not reported	60
Design	Quasi-experimental	Comparative AWE evaluation	Mixed methods (pre/post + interviews)	Survey + interviews	Scoring validity study	Corpus-based experimental	Exploratory process-tracing	Experimental
Tool	Generative AI (ChatGPT, Grammarly)	Pigai, iWrite, Awrite	ChatGPT-assisted instruction	AI-powered feedback (Pigai, Grammarly)	ChatGPT (GPT-4 mini)	DECOR (Detect-Explain-Rewrite)	Pigai AWE	AWE (Criterion)
Instruments	Writing pre/post-test, engagement questionnaire	AEE rubric scores, student survey	Pre/post writing test; motivation survey	Online survey, semi-structured interviews	IELTS rubric comparison, human vs AI scores	Human vs NLP coherence scoring	Revision log data, writing drafts	Teacher feedback logs, student drafts
Outcomes	Writing proficiency, engagement	Writing accuracy, coherence, and perceptions	Writing quality, motivation	Perceptions, benefits, challenges	Scoring accuracy, feedback quality	Coherence improvement, feedback quality	Revision behaviour, uptake	Writing quality, teacher feedback use
Effect Size/ Stats	t (98) =3.45, p<.01; dâ‰ˆ0.46	r=0.67 with human rater; p<.001	F (1,48) =5.87, p=.019; Î·Â²=0.11	Descriptive statistics	r=0.72, p<.01	r=0.81 with human scores	Descriptive; trend analysis	Î·Â²=0.21, p<.001

Criteria	Detailed Information in Appendix B							
Quality Notes	Peer-reviewed; strong alignment with EMI context	Well-reported metrics; high reliability	Frontiers in Psychology; robust qualitative triangulation	Good thematic depth; single-institution sample	Validates AI scoring reliability; limited sample	Technical NLP design; strong computational evidence	Well-reported; small sample	Strong empirical design; aligns with CALL focus
Significance Note	Significant improvement reported (N=100; Quasi-experimental; key metric: t (98) =3.45, p<.01; dâ‰ˆ0.46)	Significant improvement reported (N=178; Comparative AWE evaluation; key metric: r=0.67 with human rater; p<.001)	Significant improvement reported (N = 50; mixed-methods design [pre/post-tests and interviews]; $F(1, 48) = 5.87$, $p = .019$, $\eta^2 = 0.11$).	Effect metrics reported (Descriptive statistics); check original for p-values (N=Not reported)	Significant improvement reported (N=50 essays; Scoring validity study; key metric: r=0.72, p<.01)	Significant improvement reported (corpus-based analysis using EFCAMDAT; correlation with human scores r = .81).	Effect metrics reported (Descriptive; trend analysis); check original for p-values (N=Not reported)	Significant improvement reported (N=60; Experimental; key metric: Î·Â²=0.21, p<.001)

APPENDIX B

Tiandem-Adamou (2024)

This study examined the effectiveness of integrating generative artificial intelligence (GenAI) tools to support English as a Foreign Language (EFL) students' academic writing in a Chinese university operating within an English-medium-instruction (EMI) context. Using a mixed-methods, quasi-experimental design, the research compared an AI-assisted intervention group with a control group receiving traditional instruction across multiple writing dimensions, including grammar and syntax, organisation and coherence, vocabulary use, argumentation, mechanics, and overall writing quality.

Quantitative results indicated statistically significant improvements in the AI-assisted group relative to the control group ($p < .01$), with small-to-medium effect sizes reported across most dimensions. Notable gains were observed in grammatical accuracy, mechanics, coherence, and overall writing performance. Additional analyses showed improvements in post-intervention writing scores across different measurement scales, supporting the reliability of the findings. While overall student engagement increased modestly, specific forms of engagement—particularly peer collaboration and AI-supported personalised feedback—were more strongly associated with improvements in writing quality.

Qualitative findings revealed that students valued GenAI for its immediacy, clarity, and actionable feedback, particularly for grammar, sentence structure, and vocabulary refinement. Participants reported that AI-generated feedback enhanced motivation and supported iterative drafting and revision processes. However, students consistently emphasised that AI feedback should complement rather than replace teacher and peer feedback, which they perceived as more context-sensitive and pedagogically nuanced. Concerns were also raised regarding potential over-reliance on AI, occasional inaccuracies, and limitations in addressing cultural and rhetorical subtleties.

The study situates GenAI within a constructivist and cooperative learning framework, conceptualising AI as a scaffolding tool that supports collaborative knowledge construction and guided learning. Despite limitations related to sample homogeneity, short intervention duration, and the absence of longitudinal data, the findings suggest that AI writing analytics can effectively support automated evaluation and writing development in EMI contexts when integrated thoughtfully with human-mediated instruction.

Wang (2022)

This study investigated the effectiveness of automated essay evaluation (AEE) systems in supporting English as a Foreign Language (EFL) students' writing development in Chinese university contexts, which share key characteristics with English-medium-instruction (EMI) environments. Focusing on AI-powered writing analytics platforms such as Pigai.org, iWrite, and Awrite, the research examined students' writing performance, perceptions of automated feedback, and the reliability of AEE scores in comparison with human raters.

Quantitative analyses revealed significant improvements in students' writing quality over time, as reflected in increased AEE-generated scores and essay length across multiple writing tasks. Effect size estimates ranged from medium to large (Cohen's $d = 0.51–0.79$), indicating meaningful gains in grammatical accuracy, mechanics, syntactic complexity, and overall writing performance. Independent evaluations by experienced human raters of selected essays corroborated the automated scores, suggesting that AEE systems can provide reliable and consistent assessments of certain linguistic dimensions of writing. Improvements were also associated with enhanced learner self-monitoring and revision practices.

Student perceptions of AEE systems were generally positive, particularly regarding the immediacy, clarity, and consistency of feedback on grammar, usage, and mechanics. However, participants reported that automated feedback was less effective for discourse-level features such as organisation, logical development, and rhetorical coherence, areas in which teacher feedback was perceived as more nuanced and context-sensitive. Some technical limitations were also noted, including occasional misidentification of lexical or syntactic errors and insufficient sensitivity to semantic or pragmatic nuances.

The study highlights that learner expectations influenced satisfaction with AEE systems, with alignment between expected and perceived performance contributing to positive engagement. Drawing on expectancy–disconfirmation theory and computer-assisted language learning frameworks, the findings suggest that AEE systems are most effective when integrated with teacher guidance and peer feedback. Overall, the study indicates that AI writing analytics can serve as a valuable tool for automated evaluation and formative support in EMI-related contexts, particularly for lower-level linguistic features, while human oversight remains essential for higher-order writing skills and pedagogical interpretation.

Chen (2025)

This study examined Chinese university students' perceptions of AI-powered feedback tools—including automated writing evaluation (AWE), generative AI (GAI), and corpus-based feedback—in English writing within EFL contexts comparable to English-medium-instruction (EMI) settings. Using qualitative and survey-based methods, the research explored learners' cognitive and affective engagement with AI feedback rather than direct writing performance outcomes.

Findings indicate that students valued AI tools for providing immediate, accessible, and personalised feedback, particularly for grammatical accuracy, clarity, and surface-level language refinement. AI feedback was perceived as useful for supporting autonomous learning and alleviating teacher workload in large classes. However, students reported limitations related to vague explanations, insufficient semantic and cultural sensitivity, and restricted support for higher-order writing skills such as argumentation, coherence, and stylistic flexibility. Concerns were also raised about over-reliance on AI, potential erosion of writing identity, and algorithmic bias in interpreting non-native English usage.

To address these challenges, the study proposed a Student–Teacher–AI collaboration model, emphasising AI literacy, critical engagement with feedback, and teacher mediation. Overall, the findings suggest that AI writing analytics can support automated evaluation in EMI-related contexts when embedded within balanced pedagogical frameworks rather than used as standalone feedback mechanisms.

Sajid, Amjad, and Khan (2025)

This study investigated the use of natural language processing (NLP) techniques to enhance second-language writing assessment, with a particular focus on coherence and cohesion—dimensions that are difficult to assess consistently through human scoring. Employing the DECOR (Detect, Explain, and Rewrite) framework and the EF-Cambridge Open Language Database (EFCAMDAT), the research examined how corpus-informed AI systems can support scalable and reliable writing evaluation.

Findings indicate that NLP-driven feedback correlated well with human judgments, particularly in identifying incoherence and offering constructive revision suggestions. By moving beyond surface-level error detection, the DECOR framework demonstrated potential for formative assessment of discourse-level writing features. The study highlighted the pedagogical value of explainable feedback in promoting learner independence and iterative revision.

Although the research was not conducted explicitly in EMI settings, its implications are relevant to EMI contexts where instructors face time and scalability constraints. The study suggests that AI writing analytics can enhance automated evaluation of ESL writing by addressing coherence more effectively, illustrates the promise of corpus-informed NLP approaches, and underscores the importance of integrating AI feedback with instructional guidance.

Song and Song (2023)

This mixed-methods study examined the effectiveness of AI-assisted language learning—specifically through ChatGPT—in improving academic writing skills and motivation among Chinese EFL university students. Quantitative results from pre-test and post-test comparisons indicated significant improvements in writing performance, including organisation, coherence, grammatical accuracy, and vocabulary use, for students receiving AI-assisted instruction.

Qualitative findings showed that learners valued AI tools for their accessibility, immediacy, and personalised feedback, which supported drafting, revision, and self-regulation. Participants reported increased motivation and confidence, perceiving AI as a supportive writing aid rather than a replacement for learning. However, concerns were expressed regarding occasional contextual inaccuracies and the risk of over-reliance on AI-generated suggestions, which could limit independent writing development.

The study, grounded in social constructivist theory, conceptualised AI as a scaffolding mechanism operating within learners' zones of proximal development. Despite limitations related to sample size and lack of longitudinal data, the findings suggest that AI writing analytics can function as effective automated evaluation and instructional support tools in EMI contexts when combined with pedagogical oversight.

Uyar and Büyükahıska (2025)

This study evaluated the effectiveness of ChatGPT as an automated essay scoring and feedback tool by comparing its assessments with those of experienced human raters using IELTS Task 2 writing descriptors. The analysis focused on essays written by B2-level EFL learners across multiple genres.

Results showed that ChatGPT generated detailed and structured feedback, often providing broader commentary than human raters. However, correlations between AI-generated scores and human ratings ranged from weak to moderate, indicating variability in scoring consistency. While instructors acknowledged the usefulness of AI feedback for revision support, they emphasised that learners' ability to interpret and apply feedback critically was essential for its effectiveness.

The study highlighted concerns regarding off-task comments, scoring reliability, and ethical implications, reinforcing the need for cautious integration. Although not conducted specifically in EMI contexts, the findings are relevant to EMI assessment practices, suggesting that AI writing analytics may serve as a supplementary evaluation tool rather than a replacement for human judgment.

Link, Mehrzad, and Rahimi (2020)

This study examined the effectiveness of automated writing evaluation (AWE) tools in second-language writing, with a focus on their pedagogical value relative to teacher feedback. Findings across studies revealed mixed evidence regarding AWE's impact on writing improvement, with teacher feedback generally rated as more individualised, focused, and pedagogically meaningful.

However, AWE tools were found to support revision and redrafting by providing consistent feedback on lower-level language features, potentially allowing teachers to concentrate on higher-order writing concerns. L2 learners appeared more receptive to automated feedback when it complemented teacher input rather than replaced it.

The review also emphasised the need for stronger validation frameworks to evaluate AWE systems systematically. Overall, the findings suggest that AI writing analytics can contribute to automated evaluation and formative feedback in EMI contexts when used as part of a hybrid feedback model combining human and technological resources.

Yang, Gao, and Shen (2024)

This study explored learner engagement with AI-based automated writing evaluation (AWE) systems, particularly Pigai, among Chinese EFL learners through multiple revision cycles. The analysis focused on how students interacted with machine-generated feedback over time.

Findings indicated that AWE effectively reduced linguistic error rates and improved writing accuracy by providing timely diagnostic feedback on grammar, collocation, and mechanics. However, students engaged less consistently with higher-level linguistic resources offered by the system, and limitations were reported in addressing creativity, conceptual development, and writing style.

While the scalability and efficiency of AWE were identified as valuable in large instructional contexts comparable to EMI environments, the study emphasised that AI systems remain limited in supporting higher-order writing skills. The authors conclude that AI writing analytics are effective for automated evaluation of linguistic accuracy but should complement, not replace, instructor feedback to support nuanced writing development in EMI settings.