Available online at
http://ejournal.uin-suska.ac.id/index.php/jealt
J.E.A.L.T Vol. 16, No. 1, 2025, pp. 80 – 92
ISSN: 2407-0998 | E-ISSN 2407-1005

Leveraging ChatGPT in EFL Writing Assessment: A Systematic Literature Review

Dodi Settiawan

Postgraduate Programme,
Universitas Islam Negeri Sultan Syarif Kasim Riau,
Pekanbaru, Riau, Indonesia

dodi.settiawan@uin-suska.ac.id

ABSTRACT: This systematic literature review investigates the role and effectiveness of ChatGPT in assessing English as a Foreign Language (EFL) learners' writing performance. Guided by the PRISMA 2020 framework, 26 peer-reviewed empirical and conceptual studies published between 2022 and 2025 were analyzed. The findings reveal that ChatGPT has been leveraged across multiple assessment functions, including automated writing evaluation, corrective feedback, peer review simulation, rubric-based scoring, and prewriting support. ChatGPT demonstrates significant promise in formative assessment contexts, offering feedback comparable in quantity and breadth to human teachers, particularly on surface-level writing features such as grammar, organization, and vocabulary. However, its performance in summative assessment remains limited due to inconsistencies with human judgment and difficulties in evaluating higher-order writing skills, such as argumentation and style. Furthermore, while ChatGPT fosters learner engagement and confidence, concerns persist regarding equity, fairness, and the authenticity of feedback for diverse learner populations. The study underscores the need for rubric standardization, pedagogical integration, and longitudinal research to ensure ethical and effective implementation. Ultimately, ChatGPT is most impactful when used to complement, rather than replace, human expertise in writing assessment.

KEYWORDS: ChatGPT, EFL writing assessment, automated feedback, formative assessment

1 INTRODUCTION

Recent advances in generative artificial intelligence (AI), particularly the introduction of large language models (LLMs) such as ChatGPT, have significantly reshaped the landscape of language education. AI-powered tools now support a wide range of pedagogical applications, including content generation, language modeling, translation, and automated feedback (Huang et al., 2023). Among these, writing instruction has emerged as a central domain where generative AI demonstrates considerable promise. For English as a Foreign Language (EFL) learners, who often struggle with syntactic accuracy, lexical choice, and text organization, access to intelligent language models offers unprecedented opportunities for personalized and scalable support (Zawacki-Richter et al., 2022). ChatGPT, developed by OpenAI, is particularly notable for its conversational capabilities, linguistic flexibility, and rapid user adoption across educational contexts. Its capacity to generate coherent, contextually

*Correspondence: dodi.settiawan@uin-suska.ac.id

relevant, and stylistically varied text positions it as a powerful candidate for writing instruction and assessment.

Parallel to its growth in educational use, there has been increasing interest in the potential of ChatGPT to serve as a tool for writing assessment. Traditional approaches to EFL writing evaluation often demand intensive teacher labor and are prone to variability in feedback quality and consistency (Weigle, 2002). Automated writing evaluation (AWE) tools have long attempted to address these challenges, but many earlier systems were limited to surface-level features such as grammar and mechanics (e.g., e-rater, Pigai). In contrast, ChatGPT claims to offer a more nuanced and context-sensitive engagement with learner writing, potentially expanding the scope of automated feedback and assessment.

Despite the growing interest in ChatGPT, empirical and conceptual clarity regarding its role in EFL writing assessment remains limited. While anecdotal reports and preliminary experiments suggest that ChatGPT can function as both a feedback provider and an automated scorer, the literature lacks a systematic synthesis of its pedagogical effectiveness, contextual limitations, and practical applications. In particular, it is unclear how ChatGPT compares to traditional human feedback or other AI-based tools in formative and summative assessment contexts. Additionally, questions persist about the quality, fairness, and interpretability of ChatGPT-generated feedback, especially for learners at varying proficiency levels or from diverse cultural and educational backgrounds.

Moreover, ethical and pedagogical concerns complicate ChatGPT's integration into assessment practices. Some scholars argue that AI-generated feedback risks promoting a formulaic approach to writing, potentially discouraging learner autonomy and creativity (Kern, 2015). Others warn that overreliance on AI tools may exacerbate inequalities, as more proficient students are better equipped to interpret and apply feedback effectively (Tsai et al., 2024). From a teacher's perspective, there is apprehension regarding the potential displacement of human judgment in writing evaluation, particularly in high-stakes contexts where nuance, voice, and rhetorical intent play crucial roles. Collectively, these tensions highlight the urgent need for critical, evidence-based exploration of ChatGPT's actual and potential roles in EFL writing assessment.

In response to these emerging questions and gaps, this study presents a systematic literature review of recent research on ChatGPT's use in EFL writing assessment. Drawing on 26 peer-reviewed empirical and conceptual studies published between 2022 and 2025, the review aims to synthesize current knowledge regarding (1) the ways in which ChatGPT has been employed to assess EFL learners' writing performance, and (2) how effective it is in doing so, according to existing literature.

The review follows the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework to ensure methodological transparency and rigor. Included studies were identified through a structured search of academic databases using Boolean combinations of relevant terms (e.g., "ChatGPT", "EFL writing", "assessment", "feedback"). Only peer-reviewed journal articles that explicitly addressed ChatGPT's role in EFL writing assessment were included. Exclusions were applied to non-peer-reviewed sources, conference proceedings, and studies focusing on other AI tools.

The analysis revealed a wide spectrum of ChatGPT applications in writing assessment, including automated scoring, written corrective feedback, rubric-based evaluation, peer review simulation, and pre-writing support. The results also indicate that while ChatGPT is generally effective in delivering formative feedback and enhancing learner engagement, its reliability as a summative assessment tool remains questionable due to misalignment with

human raters and limitations in handling rhetorical nuance and genre conventions. The study concludes by identifying three critical areas for further development: rubric standardization, pedagogical integration, and equity-centered implementation.

By consolidating current evidence, this review contributes to the growing discourse on the integration of generative AI in language education and offers practical insights for educators, researchers, and policymakers seeking to balance innovation with pedagogical integrity in writing assessment.

2 METHODOLOGY

This study employed a systematic literature review methodology, guided by the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework. The review aimed to identify, screen, and synthesize existing research on how ChatGPT has been leveraged to assess English as a Foreign Language (EFL) learners' writing. This method was selected for its transparency, replicability, and comprehensiveness in synthesizing existing evidence across multiple sources. A systematic search was conducted using an academic database, ERIC, using combinations of the following keywords and Boolean operators: ("ChatGPT" OR "Generative AI") AND ("EFL writing" OR "English as a foreign language" AND "writing assessment") AND ("assessment" OR "feedback" OR "evaluation").

The search was limited to peer-reviewed journal articles published from 2022 to 2025, in English, and focused on empirical or conceptual studies related to the assessment of EFL writing using ChatGPT. The following inclusion and exclusion criteria were applied:

2.1 INCLUSION CRITERIA

a. Peer-reviewed journal articles
b. Focus on ChatGPT as a tool for assessing or supporting assessment in EFL writing
c. Articles published within the specified date range
d. Empirical studies, case studies, or theoretical papers relevant to EFL contexts

2.2 EXCLUSION CRITERIA

a. Non-peer-reviewed sources (e.g., reports, blog posts, editorials)
b. Conference proceedings
c. Studies focusing on general AI or other AWE tools (e.g., Pigai) without specific mention of ChatGPT
d. Articles unrelated to EFL learners or not focusing on writing assessment

The selection process followed the PRISMA 2020 flow, encompassing four phases: Identification, Screening, Eligibility, and Inclusion. In the identification phase, a total of 36 records were retrieved through initial database searches. After applying automatic filters for publication date and peer-review status, 3 records were removed. Then, continued to screening phase in which the remaining 33 records were screened by title and abstract. 5 articles were excluded due to one or more of the following reasons:

a. Focus on general AI or digital writing tools rather than ChatGPT
b. Discussion of alternative automated writing evaluation tools (e.g., Pigai)
c. Lack of relevance to EFL writing assessment

In the eligibility phase, 28 full-text articles were assessed. 2 articles were excluded at this stage as they were reports or conference proceedings rather than peer-reviewed journal articles. Finally, after applying all inclusion and exclusion criteria, 26 studies were selected for full analysis in the review.

The PRISMA flow diagram below summarizes the study selection process:

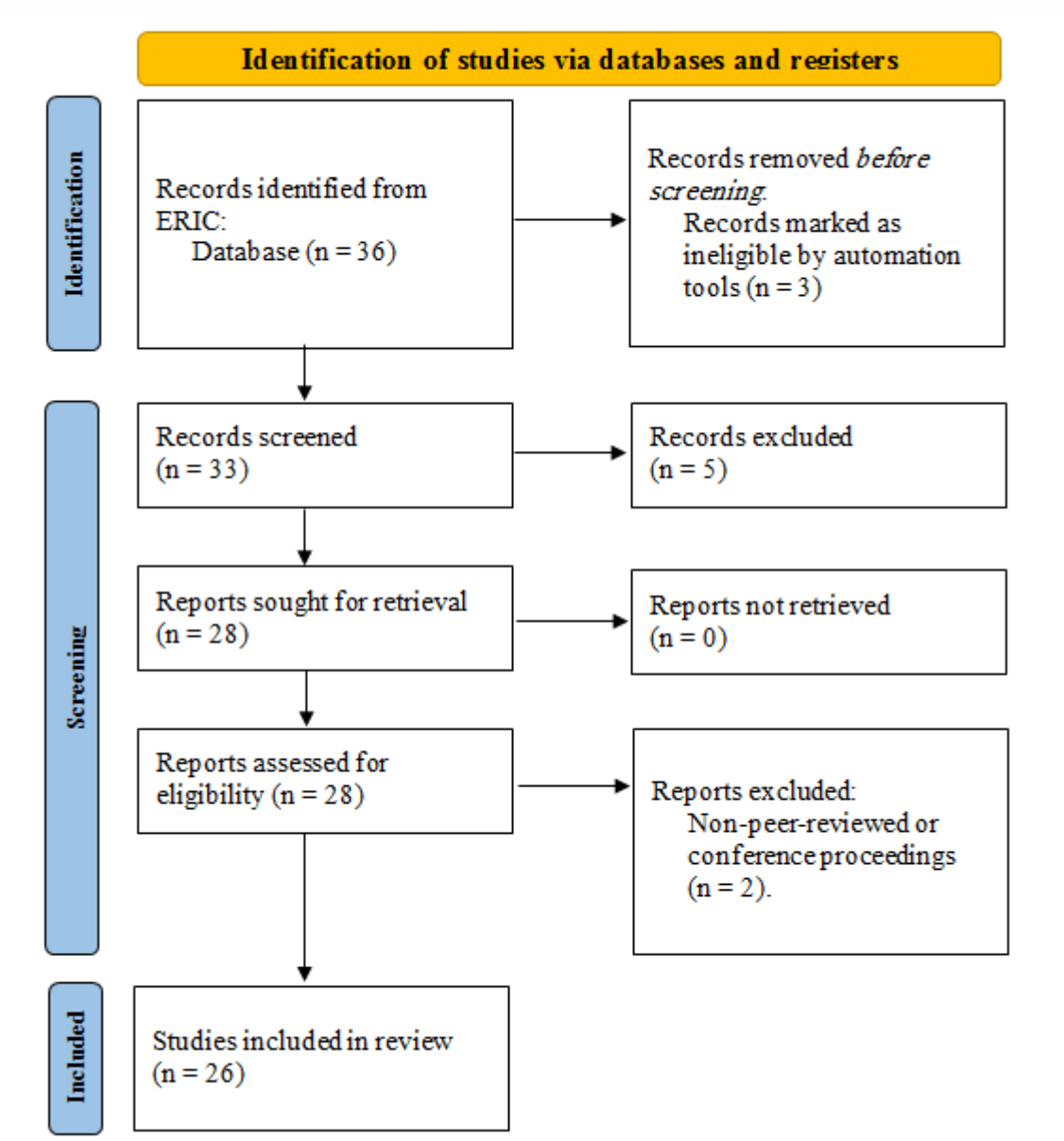

Fig. 1. Article Selection Process in PRISMA Flow Diagram

This rigorous selection process, guided by PRISMA 2020 standards, ensured the inclusion of high-quality, relevant studies for systematic review. From an initial pool of 36 records, the four-phase screening process—incorporating automated filters and manual reviews of titles, abstracts, and full texts—yielded 26 studies that met all inclusion criteria. The systematic exclusion of non-peer-reviewed sources, non-ChatGPT tools, and irrelevant studies strengthened the review's validity, while the PRISMA flow diagram (Figure 1) transparently documents this methodological rigor. This approach guarantees that the subsequent analysis is grounded in empirically sound and contextually appropriate literature.

3 RESULTS

This section presents the findings from a systematic analysis of current literature addressing two key research questions: (1) In what ways has ChatGPT been used to assess EFL learners' writing performance? and (2) How effective is ChatGPT in assessing EFL learners' writing? The results reveal diverse applications of ChatGPT in writing assessment, ranging from automated scoring and corrective feedback to collaborative peer review and rubric development. Additionally, the analysis highlights ChatGPT's effectiveness across multiple dimensions, including formative feedback quality, reliability in scoring, and learner engagement, while also identifying critical limitations. Together, these findings provide a comprehensive understanding of ChatGPT's evolving role in EFL writing assessment, its pedagogical potential, and areas requiring further refinement.

3.1 THE ROLE OF CHATGPT IN ASSESSING EFL LEARNERS' WRITING PERFORMANCE

The review of selected studies revealed that ChatGPT has been employed in diverse ways to assess EFL learners' writing performance. The findings are organized into six key themes, each highlighting distinct applications and implications of ChatGPT in writing assessment.

a. ChatGPT as an Automated Writing Evaluation (AWE) Tool

ChatGPT functions as an effective AWE tool, providing automated feedback on linguistic and structural aspects of writing, such as grammar, vocabulary, cohesion, and mechanics. Studies demonstrated its reliability in scoring and offering surface-level feedback, particularly when fine-tuned with specific rubrics. For instance, Bucol and Sangkawong (2025) found ChatGPT comparable to human raters in Thai EFL contexts, while Yavuz et al. (2025) reported high inter-rater reliability when using a five-domain rubric. However, the tool requires further refinement to address deeper writing features, such as critical thinking and creativity.

b. ChatGPT for Written Corrective Feedback (WCF)

ChatGPT has proven valuable in delivering formative feedback, particularly Written Corrective Feedback (WCF), which targets content, organization, vocabulary, and grammar. Studies by Alsofyani and Barzanji (2024) and Zeevy-Solovey (2024) highlighted ChatGPT's ability to provide fast and balanced feedback compared to traditional teacher or peer feedback. Guo and Wang (2024) further noted that ChatGPT's feedback was more comprehensive, addressing multiple dimensions of writing equally. Bai and Wei (2024) observed active learner engagement with ChatGPT's feedback, leading to noticeable revisions. These findings suggest that while ChatGPT is effective in generating WCF, its impact depends on learners' ability to interpret and apply the feedback, necessitating instructional support.

c. ChatGPT as a Peer Reviewer or Collaborative Feedback Partner

Positioning ChatGPT as a simulated peer reviewer has shown promise in reducing learner anxiety and enhancing engagement. Wang and Zhang (2024) demonstrated that ChatGPT, when used as a "digital peer," improved learners' confidence and the quality of their feedback exchanges. Similarly, Tseng and Lin (2024) integrated ChatGPT within instructional frameworks (ADDIE and TPACK) to simulate collaborative review processes, which fostered student confidence and critical thinking. These results indicate that ChatGPT can serve as a reliable and approachable alternative in contexts where peer feedback is logistically challenging.

d. ChatGPT for Essay Scoring and Standardized Rubric-Based Assessment

Several studies explored ChatGPT's ability to score essays using standardized rubrics, such as IELTS band descriptors. Uyar and Büyükahıska (2025) found statistically significant differences between ChatGPT and human raters, suggesting limitations in the tool's ability to capture nuanced language use. Conversely, Mahdi and Alkhateeb (2025) reported consistent scoring with high inter-rater reliability when using a tailored rubric. These findings underscore ChatGPT's potential for rubric-based scoring but also highlight discrepancies in its alignment with human judgment, particularly for genre-specific or culturally nuanced writing.

e. ChatGPT for Pre-Writing Planning and Draft Evaluation

Indirectly, ChatGPT has influenced writing assessment by supporting pre-writing planning and revision. Nguyen and Nguyen (2025) observed that ChatGPT enhanced learners' pre-writing strategies, leading to improved writing outcomes and affective engagement. Tsai, Lin, and Brown (2024) noted significant score improvements in revised essays after ChatGPT assistance, raising ethical concerns about fairness in evaluation. These results suggest that ChatGPT's role in formative stages introduces a new dimension to process-based assessment, where its support affects final performance metrics.

f. ChatGPT in Rubric Development and Evaluation Consistency Studies

A smaller subset of studies focused on rubric development and validation for AI-assisted writing assessment. Mahdi and Alkhateeb (2025) designed a robust rubric to evaluate AI-generated essays, while Yavuz et al. (2025) compared rubric-based evaluations across human and AI raters to validate ChatGPT's reliability. These efforts highlight the importance of standardized rubrics in ensuring transparency and consistency when integrating ChatGPT into writing assessment practices.

The synthesis of these studies reveals that ChatGPT serves multiple roles in EFL writing assessment, from automated scoring and corrective feedback to peer collaboration and rubric development. While the tool demonstrates efficiency and scalability, its effectiveness varies depending on task complexity, rubric design, and learner engagement. The table below summarizes the key use cases and their supporting evidence:

Use Case	Main Role of ChatGPT	Key Studies
Automated Writing Evaluation	Scoring essays, providing automated feedback	Bucol & Sangkawong (2025); Yavuz et al. (2025)
Written Corrective Feedback	Delivering targeted feedback	Alsofyani & Barzanji (2024); Guo & Wang (2024)
Peer Feedback Partner	Simulating peer review	Wang & Zhang (2024); Tseng & Lin (2024)
Essay Scoring (AES)	Rubric-based scoring	Uyar & Büyükahıska (2025); Mahdi & Alkhateeb (2025)
Pre-writing & Revision Support	Enhancing planning and revision	Nguyen & Nguyen (2025); Tsai et al. (2024)
Rubric Development	Validating assessment rubrics	Mahdi & Alkhateeb (2025); Yavuz et al. (2025)

In conclusion, ChatGPT offers versatile applications in EFL writing assessment, but its integration requires careful consideration of pedagogical goals, rubric standardization, and the

balance between automated and human evaluation. Future research should address its limitations in nuanced assessment and explore strategies to maximize its educational benefits.

3.2 THE EFFECTIVENESS OF CHATGPT IN ASSESSING EFL LEARNERS' WRITING PERFORMANCE

The analysis of 26 studies reveals a nuanced and context-dependent perspective on the effectiveness of ChatGPT in assessing EFL learners' writing. The findings are organized into five key themes, each addressing distinct dimensions of effectiveness, including formative feedback, automated scoring, writing improvement, learner engagement, and limitations.

a. Comparable to Human Feedback in Formative Assessment

ChatGPT demonstrates significant effectiveness in delivering formative feedback, particularly in areas such as grammar, vocabulary, and organization. Studies indicate that its feedback is comparable to that provided by human teachers. For example, Alsofyani and Barzanji (2024) found no statistically significant difference in writing improvement between students who received feedback from ChatGPT and those who received teacher feedback. Similarly, Zeevy-Solovey (2024) reported that students rated ChatGPT feedback as nearly as helpful as teacher feedback, with many preferring a combination of both. Guo and Wang (2024) further noted that ChatGPT generated a higher quantity of feedback, evenly distributed across content, organization, and language. Polakova and Ivenz (2024) observed measurable improvements in grammar, conciseness, and passive voice usage after students engaged with ChatGPT-assisted feedback. These findings suggest that ChatGPT can serve as a viable tool for formative assessment, particularly when integrated thoughtfully into the feedback cycle.

b. Moderately Reliable as an Automated Essay Scoring (AES) Tool

While ChatGPT exhibits high consistency in scoring, its alignment with human ratings varies, particularly in high-stakes assessment contexts. Uyar and Büyükahıska (2025) identified statistically significant discrepancies between ChatGPT and human raters when using IELTS rubrics, with ChatGPT often assigning lower scores. Conversely, Yavuz et al. (2025) reported very high inter-rater reliability (ICC = 0.972) when ChatGPT was fine-tuned for specific domains, though challenges remained in assessing stylistic and nuanced aspects of writing. Mahdi and Alkhateeb (2025) developed a tailored rubric and found that ChatGPT produced scores consistent with those generated by other AI tools like Claude. These results highlight ChatGPT's potential as an AES tool but underscore the need for further refinement and rubric alignment to ensure reliability in summative assessment contexts.

c. Effective in Promoting Writing Improvement through Feedback Integration

ChatGPT has proven effective in supporting learners' ability to revise and enhance their writing quality. Bai and Wei (2024) found that students actively integrated ChatGPT's reformulations into their revisions, with the quality of their noticing behavior influencing the extent of uptake. Tsai, Lin, and Brown (2024) observed significant improvements in vocabulary, grammar, content, and organization in essays revised with ChatGPT assistance. Nguyen and Nguyen (2025) further demonstrated that ChatGPT's use during prewriting stages led to better planning strategies and overall writing quality. These findings suggest that ChatGPT is particularly effective in revision-based assessment, especially when paired with metacognitive strategies that encourage learners to critically engage with feedback.

d. Positive Impact on Learner Engagement and Writing Confidence

ChatGPT contributes to enhanced emotional and behavioral engagement, motivation, and self-efficacy among EFL learners. Teng and Huang (2025) reported improvements in affective and behavioral engagement, though cognitive gains were less pronounced. Teng (2024) emphasized the role of metacognitive awareness in maximizing the effectiveness of ChatGPT feedback. Mohammed and Khalid (2025) noted that ChatGPT feedback boosted learners' motivation, peace of mind, and writing proficiency, while Lai (2025) documented gains in resilience and linguistic accuracy. These outcomes indicate that ChatGPT's formative assessment capabilities are particularly beneficial for learners with lower confidence or autonomy, fostering a supportive environment for writing development.

e. Ethical and Practical Limitations

Despite its advantages, ChatGPT's effectiveness is tempered by several limitations. Uyar and Büyükahıska (2025) cautioned that ChatGPT has not yet achieved sufficient proficiency for practical AES use in high-stakes settings. Tsai, Lin, and Brown (2024) raised fairness concerns, noting that weaker students improved disproportionately, potentially skewing grade distributions. Alsalem (2024) highlighted teacher skepticism about replacing human judgment with AI, advocating for cautious integration. Won et al. (2025) identified challenges in replicating authentic learner language, as ChatGPT struggled to generate formulaic patterns typical of L2 learners. These limitations underscore the importance of context-sensitive implementation and the need for ongoing research to address ethical and practical challenges.

The synthesis of these studies demonstrates that ChatGPT is moderately to highly effective in assessing EFL writing, with its strengths lying in formative feedback, revision support, and learner engagement. However, its reliability in automated scoring remains inconsistent, and ethical concerns necessitate careful consideration. The table below summarizes the key dimensions of effectiveness and their supporting evidence:

Dimension	**Effectiveness of ChatGPT**	**Key Studies**
Formative Feedback	Comparable to teacher feedback in scope and usefulness	Alsofyani & Barzanji (2024); Zeevy-Solovey (2024)
Automated Scoring (AES)	Consistent but not fully aligned with human raters	Uyar & Büyükahıska (2025); Yavuz et al. (2025)
Revision and Writing Quality	Supports error noticing and effective revision	Bai & Wei (2024); Tsai et al. (2024)
Learner Engagement	Boosts confidence, motivation, and self-efficacy	Teng (2024); Mohammed & Khalid (2025)
Limitations	Issues with fairness, reliability, and authenticity	Tsai et al. (2024); Won et al. (2025)

In conclusion, ChatGPT is a promising tool for formative writing assessment in EFL contexts, offering scalable feedback and enhancing learner engagement. However, its role in high-stakes automated scoring requires further development to address reliability and

alignment with human judgment. Ethical considerations, such as fairness and authenticity, must also be prioritized to ensure equitable assessment practices. Future research should explore strategies to mitigate these limitations while maximizing ChatGPT's pedagogical potential.

4 DISCUSSION

The findings of this study illuminate both the multifaceted applications and the nuanced effectiveness of ChatGPT in assessing EFL learners' writing. By synthesizing evidence from 26 studies, this discussion contextualizes the results within broader pedagogical and technological frameworks, addressing implications, limitations, and future directions.

4.1 THE DUAL ROLE OF CHATGPT: TOOL AND PARTNER IN WRITING ASSESSMENT

ChatGPT's versatility is evident in its dual capacity as an automated assessment tool and a collaborative feedback partner. As an AWE tool, it demonstrates reliability in evaluating surface-level features (e.g., grammar, vocabulary) but struggles with higher-order skills like critical thinking (Bucol & Sangkawong, 2025; Yavuz et al., 2025). This aligns with prior research on AWE systems (e.g., Warschauer & Ware, 2006), which notes that while AI excels in efficiency, it lacks the contextual and cultural awareness of human raters.

Conversely, ChatGPT's role as a peer reviewer (Wang & Zhang, 2024) or formative feedback provider (Guo & Wang, 2024) introduces a paradigm shift in learner engagement. By reducing anxiety and offering immediate, balanced feedback, it addresses gaps in traditional peer review systems, particularly in large or remote classrooms. However, its effectiveness hinges on instructional scaffolding—learners who lack metacognitive strategies may misinterpret or overlook feedback (Bai & Wei, 2024). This underscores Vygotskian principles of guided learning (1978), suggesting ChatGPT is most impactful when integrated into a structured pedagogical framework.

4.2 EFFECTIVENESS: STRENGTHS AND CONTEXTUAL LIMITATIONS

a. Formative Feedback vs. Summative Scoring

ChatGPT's efficacy varies markedly between formative and summative contexts. For formative assessment, it rivals human teachers in feedback quantity and breadth (Alsofyani & Barzanji, 2024), yet its quality depends on task complexity. For instance, while it improves grammar and organization (Polakova & Ivenz, 2024), its feedback on argumentation or style remains generic (Uyar & Büyükahıska, 2025).

In summative contexts, ChatGPT's inconsistency with human raters (Mahdi & Alkhateeb, 2025) mirrors challenges seen in earlier AES tools like e-rater (Attali, 2013). Discrepancies arise in culturally nuanced tasks (e.g., IELTS writing prompts), where ChatGPT may undervalue idiomaticity or rhetorical flair. This suggests that while rubric fine-tuning enhances reliability (Yavuz et al., 2025), AI cannot yet replicate the holistic judgment of human evaluators.

b. Learner Engagement and Equity Concerns

The positive impact on motivation and self-efficacy (Teng, 2024; Mohammed & Khalid, 2025) highlights ChatGPT's potential to democratize feedback access, especially for underserved learners. However, Tsai et al. (2024) raise critical equity issues: weaker students may over-rely on ChatGPT, inflating grades without commensurate skill

development. This echoes concerns about "techne" overshadowing "praxis" in digital learning (Kern, 2015), where tool dependence undermines autonomous learning.

The existing literature highlights three critical challenges associated with the use of ChatGPT in EFL writing assessment. First, concerns about fairness emerge as ChatGPT's feedback may unintentionally exacerbate achievement disparities, particularly when learners' access to the tool or their ability to interpret its feedback varies by proficiency level (Alsalem, 2024). Second, issues of authenticity arise due to ChatGPT's current limitations in accurately replicating the linguistic features typical of second language (L2) learners, potentially resulting in a misalignment with the realities of learner writing (Won et al., 2025). Third, the evolving role of educators remains a point of contention; while ChatGPT offers scalable support, teachers express skepticism about its capacity to replace human judgment, instead advocating for its use as a supplementary tool within hybrid assessment models (Alsalem, 2024).

To maximize the pedagogical potential of ChatGPT, several strategic directions warrant attention. Rubric standardization is essential, particularly the development of AI-sensitive evaluation frameworks that integrate both surface-level linguistic accuracy and higher-order cognitive skills (Mahdi & Alkhateeb, 2025). Pedagogical integration should also be prioritized, including the design of instructional models—such as those grounded in the TPACK framework—that scaffold effective and ethical use of generative AI tools in classroom contexts (Tseng & Lin, 2024). Finally, longitudinal research is needed to examine the sustained impact of ChatGPT on EFL learners' writing proficiency, autonomy, and engagement over time.

ChatGPT represents a significant advancement in EFL writing assessment, offering scalable feedback and fostering engagement. Yet, its effectiveness is bounded by task type, learner context, and ethical considerations. As the field evolves, a collaborative approach—where AI complements human expertise—will be essential to balance innovation with pedagogical integrity. Future research should explore adaptive feedback models and equity-focused implementations to ensure ChatGPT serves as a bridge, not a barrier, to writing development.

5 CONCLUSION

This systematic review has demonstrated that ChatGPT holds significant potential as a multifaceted tool for assessing EFL writing, while simultaneously revealing important limitations that must be addressed. The analysis of 26 empirical studies indicates that ChatGPT serves effectively as both an automated assessment tool and collaborative feedback partner, particularly in formative contexts where it demonstrates comparable performance to human instructors in providing grammatical and organizational feedback. However, its reliability diminishes in summative assessment scenarios and when evaluating higher-order writing skills, underscoring the continued necessity of human judgment in nuanced evaluation contexts.

The findings highlight three critical considerations for successful implementation. First, ChatGPT's pedagogical value is most evident when integrated thoughtfully within existing instructional frameworks, where it can supplement rather than replace educator expertise. Second, the tool's current limitations in cultural sensitivity and linguistic nuance necessitate careful rubric development and continuous refinement of assessment criteria. Third, ethical

concerns regarding equity of access and potential over-reliance by learners must be proactively addressed through institutional policies and teacher training programs.

Future research should prioritize longitudinal studies examining ChatGPT's long-term impact on writing proficiency, investigations into adaptive feedback models for diverse learner populations, and development of comprehensive training programs for educators. As AI writing assessment tools continue to evolve, maintaining a balanced approach that leverages technological advantages while preserving essential human elements of language instruction and evaluation will be paramount. The successful integration of ChatGPT into EFL writing assessment ultimately depends on establishing collaborative systems that combine AI efficiency with pedagogical expertise, ensuring both the quality of assessment outcomes and the development of authentic writing competencies.

REFERENCES

Alsalem, M. S. (2024). EFL teachers' perceptions of the use of an AI grading tool (CoGrader) in English writing assessment at Saudi universities: an Activity Theory Perspective. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2430865

Alsofyani, A. H., & Barzanji, A. M. (2024). The Effects of ChatGPT-Generated Feedback on Saudi EFL Learners' Writing Skills and Perception at the Tertiary Level: A Mixed-Methods Study. Journal of Educational Computing Research, 63(2), 431-463. https://doi.org/10.1177/07356331241307297 (Original work published 2025)

Amani, N. & Bisriyah, M. (2025). University Students' Perceptions of AI-Assisted Writing Tools in Supporting Self-Regulated Writing Practices. *IJELTAL*, Vol 10, No 1, http://dx.doi.org/10.21093/ijeltal.v10i1.1942

Attali, Y. (2013). Validity and reliability of automated essay scoring. *ETS Research Report Series.*

Aydın, S. & Zeinolabedini, M. (2024). Integrating Artificial Intelligence into Foreign Language Learning: Learners' Perspectives. *EJER Congress 2024 International Eurasian Educational Research Congress Conference Proceedings," Ani Publishing, pp. 121-126*

Bai, L. and Wei, Y. (2024). "Exploring EFL Learners' Integration and Perceptions of ChatGPT's Text Revisions: A Three-Stage Writing Task Study," in IEEE Transactions on Learning Technologies, vol. 17, pp. 2161-2172, , doi: 10.1109/TLT.2024.3491864

Bucol, J. L., & Sangkawong, N. (2024). Exploring ChatGPT as a writing assessment tool. Innovations in Education and Teaching International, 62(3), 867–882. https://doi.org/10.1080/14703297.2024.2363901

Bucol, J., & Sangkawong, N. (2025). Exploring ChatGPT as a writing assessment tool. Innovations in Education and Teaching International, 62(3) 867-882. doi:10.1080/14703297.2024.2363901

Gozali, I., Wijaya, A. R. T., Lie, A., Cahyono, B. Y., & Suryati, N. (2024). Leveraging the potential of ChatGPT as an automated writing evaluation (AWE) tool: Students' feedback literacy development and AWE tools integration framework. *The JALT CALL Journal*, *20*(1), 1–22. https://doi.org/10.29140/jaltcall.v20n1.1200

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Educ Inf Technol* 29, 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

He, Y. (2024). The Metaphor of AI in Writing in English: A Reflection on EFL Learners' Motivation to Write, Enjoyment of Writing, Academic Buoyancy, and Academic Success in Writing. The International Review of Research in Open and Distributed Learning, 25(3), 271–286. https://doi.org/10.19173/irrodl.v25i3.7769

Kern, R. (2015). *Language, literacy, and technology*. Cambridge University Press.

Lai, Z. C. (2025). The Impact of AI-Assisted Blended Learning on Writing Efficacy and Resilience. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT), 15*(1), 1-21. https://doi.org/10.4018/IJCALLT.377174

Mahdi, H. S. & Alkhateeb, A. (2025). Revolutionising Essay Evaluation: A Cutting-Edge Rubric for AI-Assisted Writing. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT), 15*(1), 1-19. https://doi.org/10.4018/IJCALLT.368226

Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: a study on proficiency, revision frequency and writing quality. *Discov Educ* **4**, 170. https://doi.org/10.1007/s44217-025-00602-7

Mohammed, S.J., & Khalid, M.W. (2025). Under the world of AI-generated feedback on writing: mirroring motivation, foreign language peace of mind, trait emotional intelligence, and writing development. *Lang Test Asia* 15, 7. https://doi.org/10.1186/s40468-025-00343-2

Nguyen, L.Q., Le, H.V. & Nguyen, P.T. (2025). A mixed-methods study on the use of chatgpt in the pre-writing stage: EFL learners' utilization patterns, affective engagement, and writing performance. Educ Inf Technol 30, 10511–10534. https://doi.org/10.1007/s10639-024-13231-8

Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education*, 11(1). https://doi.org/10.1080/2331186X.2024.2410101

Teng, M. F. & Huang, J. (2025). Incorporating ChatGPT for EFL Writing and Its Effects on Writing Engagement. International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT), 15(1), 1-21. https://doi.org/10.4018/IJCALLT.367874

Teng, M. F. (2024). Metacognitive Awareness and EFL Learners' Perceptions and Experiences in Utilising ChatGPT for Writing Feedback. European Journal of Education Research, Development and Policy. https://doi.org/10.1111/ejed.12811

Tsai, CY., Lin, YT. & Brown, I.K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. *Educ Inf Technol* **29**, 22427–22445 https://doi.org/10.1007/s10639-024-12722-y

Tseng, Y. C. & Lin, Y. H. (2024). Enhancing English as a Foreign Language (EFL) Learners' Writing with ChatGPT: A University-Level Course Design. *EJEL*, Vol. 22 No. 2. https://doi.org/10.34190/ejel.21.5.3329

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

Vygotsky, L. S. (1978). *Mind in society*. Harvard University Press.

Waked, A., Ashraf, M. W., AbdelSalam, H.; El Alaoui, K., & Pilotti, M. (2024). Success Through Error: Using Error Analysis of ChatGPT Output in English as a Foreign Language Learner Writing Instruction. In M. Shelley & O. T. Ozturk (Eds.), Proceedings of ICRES 2024-- International Conference on Research in Education and Science (pp. 320-332)

Wang, X., and Zhang, W. (2024). Generative AI as a Collaborative Companion: Enhancing Peer Feedback in EFL Writing Classes. Education Research and Perspectives, vol. 51, pp. 1-2-123.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*.

Won, D. O., Shin, Y. K., Kim, H. J., & Yoo, I. W. (2025). Advancing Language Assessment with GPT: Is It Nonnative-Language Friendly? *Language Assessment Quarterly*, *22*(1), 1–28. https://doi.org/10.1080/15434303.2024.2444349

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, 56, 150–166. https://doi.org/10.1111/bjet.13494

Zeevy-Solovey, O. (2024). Comparing peer, ChatGPT, and teacher corrective feedback in EFL writing: Students' perceptions and preferences. Technology in Language Teaching & Learning, 6(3), 1482. https://doi.org/10.29140/tltl.v6n3.1482

Zhang, X., and Umeanowai, K.O. (2025). Exploring the transformative influence of artificial intelligence in EFL context: A comprehensive bibliometric analysis. Educ Inf Technol 30, 3183–3198. https://doi.org/10.1007/s10639-024-12937-z

Zhao, D. (2025). The impact of AI-enhanced natural language processing tools on writing proficiency: an analysis of language precision, content summarization, and creative writing facilitation. *Educ Inf Technol* **30**, 8055–8086. https://doi.org/10.1007/s10639-024-13145-5

Zheldibayeva, R. (2025). GenAI as a Learning Buddy for Non-English Majors: Effects on Listening and Writing Performance. *Educational Process: International Journal*, 14, e2025051. https://doi.org/10.22521/edupij.2025.14.51