International Multidisciplinary Journal of Emerging Technologies and Applications (IMJETA)
Vol. 1, No. 5, pp. 1-20, August 2026
Received July 08, 2026; Revised July 17, 2026; Accepted August 06, 2026
Published August 31, 2026

Artificial Intelligence Tools in EFL/ESL Academic Writing Instruction: A Systematic Literature Review (2019–2025)

Rani Santika[1][0000-0003-1395-8974] and Rafi Farizki[2][0000-0001-9836-4651]
[1]Universitas Muhammadiyah Prof Dr Hamka, Indonesia
[2]Politeknik Siber Cerdika Internasional, Indonesia
rsantika851@gmail.com, rafifari90@gmail.com

Abstract. Artificial intelligence (AI) tools are increasingly used in English as a foreign or second language (EFL/ESL) academic writing classrooms, yet evidence on their instructional use and effects remains fragmented across tools, contexts, and study designs. This systematic literature review synthesized primary empirical studies published between 2019 and 2025 and retrieved from Scopus, Web of Science, ERIC, and Google Scholar, following the PRISMA 2020 guidelines, and appraised the methodological quality of the included studies. Three research questions were addressed, concerning the AI tools used in EFL/ESL writing instruction, their effects on writing quality and skills, and the associated challenges and ethical considerations. The findings indicate that three categories of tools, automated writing evaluation and grammar-checking tools, paraphrasing tools, and generative AI are used across all stages of the writing process, frequently in combination. Their reported effects reveal a tension between consistently documented gains at the surface level of grammar, mechanics, and vocabulary and less certain gains in higher-order aspects such as content, organization, and coherence, which appear mainly in studies with stronger designs or newer generative models. Recurring concerns include student overreliance, academic integrity, and the uneven distribution of benefits across learners of differing proficiency. The review contributes an EFL-specific synthesis spanning multiple tool types and educational levels, and argues that the value of AI tools depends less on the technology itself than on how it is pedagogically mediated. Implications for teacher-guided, critically evaluated use and directions for future research are discussed.

Keywords: Academic Writing; Artificial Intelligence; Second Language Writing; Systematic Review; Writing Instruction.

1. Introduction

Academic writing is among the most cognitively demanding competencies that learners of English as a foreign or second language (EFL/ESL) must acquire. Producing an effective academic text requires the simultaneous control of grammar, vocabulary, and mechanics together with higher-order concerns such as cohesion,

rhetorical organization, argumentation, and the generation and structuring of ideas (Hyland, 2019). For EFL/ESL learners, these demands are intensified by limited linguistic resources such as grammar, vocabulary, and syntax (De Wilde, 2023; Vasylets & Marín, 2022), difficulty in generating and organizing ideas in a second language (Hayes & Flower, 2016), and low writing self-confidence (Sun & Wang, 2020), making writing a persistent source of difficulty across secondary, undergraduate, and postgraduate levels.

Given these challenges, the integration of AI tools into EFL/ESL writing instruction is directly relevant to Sustainable Development Goal 4 (SDG 4): Quality Education (United Nations, 2015). Specifically, the present review aligns with three SDG 4 targets. First, it relates to Target 4.3 (equal access to affordable technical, vocational, and higher education), as it examines how scalable, often freely available digital tools can assist language learners at secondary, undergraduate, and postgraduate levels. Second, it addresses Target 4.4 (increasing the number of people with relevant skills for financial success), since the academic writing competencies supported by these tools are critical for higher education advancement and professional communication. Third, it speaks to Target 4.5 (eliminating discrimination in education), as the review explicitly considers how these tools differentially affect learners of varying proficiency levels, a question that bears directly on equity in classroom dynamics and digital resource distribution. Situating this review within the SDG 4 framework underscores the broader educational significance of understanding how AI tools can be leveraged to promote inclusive, quality writing instruction for diverse EFL/ESL learner populations.

Over the past several years, artificial intelligence (AI) tools have moved rapidly from the margins to the mainstream of writing instruction. These tools span several categories: automated writing evaluation (AWE) and grammar-checking systems such as Grammarly, whose usefulness for EFL writing was documented well before the generative turn (Ghufron & Rosyida, 2018; Barrot, 2022) and has continued to be reported since (Vu, 2024); AI-powered paraphrasing tools such as QuillBot; and, following the public release of ChatGPT in late 2022, general-purpose generative AI systems built on large language models (LLMs). Collectively, such tools offer learners immediate, individualized feedback and text-generation support at a scale and speed that traditional teacher feedback alone cannot match (Barrot, 2023). Their rapid, largely unregulated uptake in classrooms has, in turn, raised urgent questions about how they should be integrated into writing pedagogy.

A rapidly expanding body of empirical research has examined how these tools affect EFL/ESL writing, yet the evidence remains far from settled. A recurrent tension concerns the *depth* of AI's contribution. A number of studies and reviews conclude that these tools—particularly AWE systems—improve mainly surface-level features such as grammatical accuracy, spelling, and mechanics, while offering limited support for higher-order concerns such as organization, coherence, and idea development (Alsaedi, 2024; Teng, 2024; Wijaya, 2025). Other studies, frequently

involving controlled designs or newer generative models, report gains that extend to higher-order dimensions, including coherence, discursive organization, argumentation, and lexical richness (Wei et al., 2023; Mekheimer, 2025; Sanz-Tejeda et al., 2025; Chanpradit, 2025). Alongside these reported benefits, researchers consistently raise concerns about student overreliance, reduced critical engagement and metacognition, and academic-integrity issues surrounding plagiarism, authorship, and fabricated citations (Gabay et al., 2026).

Several literature reviews have begun to consolidate this evidence, but each is bounded in ways that leave an important gap. Some are tool-specific, synthesizing research on a single tool—ChatGPT (Alsaedi, 2024; Teng, 2024) or Grammarly (Wijaya, 2025)—and therefore cannot capture how the wider ecosystem of AI writing tools is used together in instruction. Others are broad in scope but not focused on language learning: recent systematic and scoping reviews synthesize AI writing tools across higher education in general (Shabir, 2025; Gabay et al., 2026) or across reading and writing in the social sciences (Sanz-Tejeda et al., 2025), while the closest predecessor concentrates on generative AI in academic writing with an emphasis on English language teaching (Chanpradit, 2025).

Across this body of reviews, four limitations recur. First, most concentrate on generative AI, giving comparatively little attention to the AWE and paraphrasing tools that remain widely used in EFL writing classrooms. Second, they are largely confined to higher education, overlooking the secondary and K-12 EFL contexts in which AI writing tools are increasingly deployed. Third, their coverage is typically restricted to the post-ChatGPT period (2023 onward), excluding the earlier AWE research that established much of what is currently known about automated feedback. Fourth, few combine an explicitly instructional focus with a critical appraisal of the methodological quality of the primary evidence, leaving readers unable to distinguish robust findings from those resting on small, self-report studies.

The present systematic literature review addresses these gaps by synthesizing empirical research published between 2019 and 2025 on the use of AI tools in EFL/ESL academic writing instruction. In contrast to prior reviews, it spans the full spectrum of tool types (AWE, paraphrasing, and generative AI) and educational levels (secondary through postgraduate), and it appraises the methodological quality of the included studies rather than treating all evidence as equivalent. Guided by the PRISMA framework, the review is organized around three research questions:

> RQ1. Which AI tools are used in EFL/ESL academic writing instruction, and how are they used?
>
> RQ2. What are the reported effects of these tools on students' writing quality and writing skills?

RQ3. What challenges, limitations, and ethical considerations are associated with their use?

By integrating evidence across tools, levels, and contexts—and by weighing the strength of that evidence—this review aims to provide teachers, researchers, and policymakers with a clearer and more differentiated understanding of where AI tools reliably support EFL/ESL writing development and where significant risks and uncertainties remain.

2. Method

2.1. Research Design

This study adopted a systematic literature review (SLR) methodology guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021). The PRISMA framework was selected because it offers a transparent, structured, and replicable procedure for identifying, screening, selecting, and synthesizing relevant studies, thereby reducing selection bias and enhancing the trustworthiness of the review. The review was designed to answer the three research questions concerning the AI tools used in EFL/ESL academic writing instruction (RQ1), their reported effects on students' writing quality and skills (RQ2), and the challenges, limitations, and ethical considerations associated with their use (RQ3).

2.2. Search Strategy

Literature was retrieved from four databases: Scopus, Web of Science, ERIC, and Google Scholar. Scopus and Web of Science were selected for their broad indexed coverage of peer-reviewed journals, ERIC was included for its education-specific indexing, and Google Scholar was added to capture additional sources, including journals indexed in Indonesian databases (e.g., SINTA), given the strong representation of Indonesian research in this field. The search was conducted in January 2026 and covered publications from January 2019 to December 2025. This window was chosen deliberately: it captures the earlier period of automated writing evaluation (AWE) and grammar-checking tools (e.g., Grammarly) as well as the rapid expansion of generative AI following the public release of ChatGPT in late 2022, allowing the review to trace developments across both eras.

The following Boolean search string was used and adapted to the syntax of each database:

("artificial intelligence" OR "AI" OR "ChatGPT" OR "generative AI" OR "large language model" OR "Grammarly" OR "QuillBot" OR "automated writing evaluation" OR "AWE") AND ("academic writing" OR "writing skill*" OR "writing instruction" OR "essay writing" OR "paraphras*") AND ("EFL" OR "ESL" OR "English as a foreign language" OR "English as a second language" OR "second language writing" OR "L2 writing")

(Note: the asterisk denotes truncation, capturing variants such as "writing skills"/"writing skill" and "paraphrasing"/"paraphrase". Field tags — title/abstract/keywords — should be adjusted per database.)

2.3. Inclusion and Exclusion Criteria

Studies were included or excluded against the predefined criteria presented in Table 1.

Table 1. Inclusion and exclusion criteria

Dimension	Inclusion	Exclusion
Focus	AI/digital writing tools used in EFL/ESL academic or English writing	Tools used outside writing (e.g., speaking, reading); general GenAI-in-education not focused on writing
Context/population	EFL/ESL learners (secondary, undergraduate, or postgraduate) and/or their teachers	First-language (L1) English writers; general higher-education populations not framed as language learners; non-language disciplines
Topic coverage	Reports tool use (RQ1), effects on writing (RQ2), and/or challenges/ethics (RQ3)	Does not address any research question
Study type	Primary empirical studies (qualitative, quantitative, or mixed methods)	Secondary studies (reviews, meta-analyses); purely conceptual papers; tech reviews without data
Publication type	Peer-reviewed journal articles and prestigious peer-reviewed conference proceedings	Grey literature (theses, dissertations); non-peer-reviewed preprints; blogs, editorials, opinion pieces
Time	Published January 2019 – December 2025	Published before 2019
Language	Published in English	Not published in English
Availability	Full text accessible	Full text unavailable

2.4. Study Selection

The selection process followed the four PRISMA phases (identification, screening, eligibility, and inclusion) and is summarized in the PRISMA flow diagram (Figure 1).

- **Identification.** The database searches returned 462 records. After 138 duplicates were removed, 324 records remained.
- **Screening.** The titles and abstracts of the 324 records were screened against the inclusion and exclusion criteria; 271 were excluded, leaving 53 for full-text assessment.
- **Eligibility.** The full texts of 53 articles were assessed; 35 were excluded, with reasons recorded (e.g., not EFL/ESL, not writing-focused, not primary empirical, grey literature, or outside the time window).
- **Inclusion.** 18 studies met all criteria and were included in the final synthesis. Screening was performed by two independent reviewers; disagreements were resolved through discussion until consensus was reached.

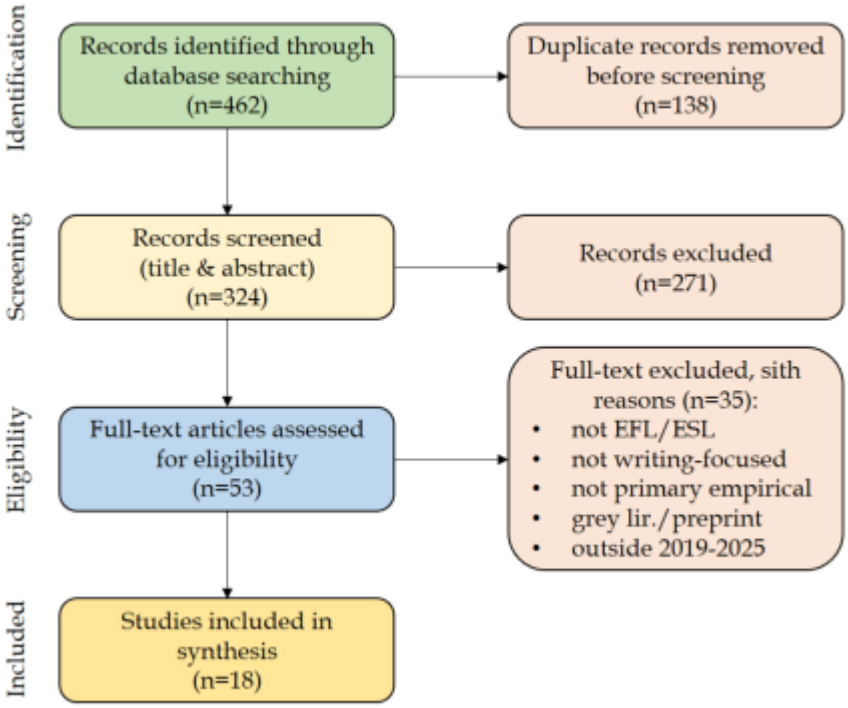

Figure 1. PRISMA 2020 flow diagram of the study selection process.

2.5. Data Extraction

A structured data-extraction form was used to code each included study consistently. The following information was extracted: (a) bibliographic details (author, year, country, journal); (b) study design and sample (participants, level, method); (c) the AI tool(s) used and how they were integrated into writing instruction (RQ1); (d) the reported effects on students' writing quality and writing skills (RQ2); and (e) the challenges, limitations, and ethical considerations reported (RQ3). Extracted data were tabulated to enable systematic comparison across studies.

2.6. Quality Appraisal

The methodological quality of the included studies was appraised using the Mixed Methods Appraisal Tool (MMAT), version 2018 (Hong et al., 2018), which provides parallel criteria for qualitative, quantitative, and mixed-methods designs and is therefore well suited to the heterogeneous corpus. Appraisal focused on the appropriateness of the design, sampling, and data collection, and on the coherence between data and findings. In synthesizing the evidence, greater interpretive weight was given to studies with stronger designs (e.g., experimental and quasi-experimental studies with measured writing outcomes) than to small-sample, self-report perception studies, so that conclusions about the effects of AI tools reflect the strength of the underlying evidence rather than the sheer number of studies.

2.7. Data Synthesis

Given the methodological heterogeneity of the included studies, a narrative and thematic synthesis (Braun & Clarke, 2006) was conducted rather than a meta-analysis. Findings were organized around the three research questions: the types and instructional uses of AI tools (RQ1), their effects on writing quality and skills (RQ2), and the associated challenges, limitations, and ethical considerations (RQ3). Within each research question, recurring patterns were grouped into themes, and convergences and divergences across studies — including the tension between surface-level and higher-order effects — were examined.

3. Results

3.1. Characteristics of the Included Studies

The review included 18 primary empirical studies published between 2019 and 2025, with a marked concentration from 2023 onward that mirrors the rapid uptake of generative AI following the public release of ChatGPT. The included studies represented a range of geographic contexts, with a strong presence of research from the Global South and, in particular, Indonesia (e.g., Rahayu et al., 2025; Hayadi et al., 2025; Nisa et al., 2026; Hamid et al., 2025; Dzahabiyyah, 2025; Werdiningsih et al., 2024; Marzuki et al., 2023; Kurniati & Fithriani, 2022), alongside studies from Saudi Arabia (Alsadoon, 2026; Mohammad et al., 2023), Egypt (Mekheimer, 2025), Ethiopia (Wale & Kassahun, 2024), Malaysia (Yang et al., 2025), Kuwait (Alghasab, 2025), the Czech Republic (Polakova & Ivenz, 2024), China (Wei et al., 2023), and South Korea (Myung et al., 2026).

Most studies were situated at the undergraduate and postgraduate levels, with a smaller number addressing secondary and K–12 EFL contexts (Alghasab, 2025; Myung et al., 2026). In terms of research design, the corpus combined studies with measured writing outcomes — including a randomized controlled trial (Wei et al., 2023), a randomized/experimental design (Mekheimer, 2025), quasi-experimental designs (Wale & Kassahun, 2024; Yang et al., 2025; Mohammad et al., 2023; Polakova & Ivenz, 2024), and a six-week classroom deployment study (Myung et al., 2026) — with a larger group of qualitative and mixed-methods studies capturing learner and teacher perceptions (e.g., Hayadi et al., 2025; Nisa et al., 2026; Hamid et al., 2025; Dzahabiyyah, 2025; Kurniati & Fithriani, 2022; Marzuki et al., 2023; Rahayu et al., 2025; Alsadoon, 2026). This distribution is important for interpretation: perception studies were more numerous, but the smaller set of experimental studies carried greater evidentiary weight for claims about actual effects on writing.

The characteristics of the 18 included studies are summarized in Table 2, and their distribution by publication year, tool category, country, and research design is shown in Figure 2.

Table 2. Characteristics of the included studies (n = 18).

No	Author	Country	Level	Design	Tool(s)	Focus
1	Polakova & Ivenz (2024)	Czech Rep.	Undergraduate	Quasi-exp.	ChatGPT	Writing outcomes (weak design)
2	Rahayu et al. (2025)	Indonesia	Undergraduate	Mixed	ChatGPT	Acceptance & perception
3	Yang et al. (2025)	Malaysia	Undergraduate	Quasi-exp.	ChatGPT	Writing outcomes + self-efficacy
4	Alsadoon (2026)	Saudi Arabia	Undergraduate	Mixed	ChatGPT	Perception + ethics
5	Wale & Kassahun (2024)	Ethiopia	Undergraduate	Quasi-exp.	Writerly + Google Docs	Writing outcomes
6	Mekheimer (2025)	Egypt	Postgraduate	Experimental	Grammarly	Writing outcomes
7	Alghasab (2025)	Kuwait	Secondary/K-12	Mixed	Grammarly, ChatGPT, Google Translate	Practices & perception (K-12)
8	Wei et al. (2023)	China	Undergraduate	Experimental	AWE / Grammarly	Writing outcomes
9	Hayadi et al. (2025)	Indonesia	Undergraduate	Qualitative	ChatGPT	Perception
10	Nisa et al. (2026)	Indonesia	Undergraduate	Qualitative	QuillBot	Practices & perception
11	Werdiningsih et al. (2024)	Indonesia	Postgraduate	Qualitative	ChatGPT	Strategic practices
12	Indria et al. (2026)	Indonesia	Undergraduate	Qualitative	QuillBot	Practices & perception
13	Hamid et al. (2025)	Indonesia	Undergraduate	Qualitative	QuillBot	Perception
14	Dzahabiyyah (2025)	Indonesia	Undergraduate	Qualitative	ChatGPT	Perception
15	Myung et al. (2026)	South Korea	Secondary/K-12	Mixed	WriteAid (custom LLM)	Interaction & outcomes (deployment)
16	Marzuki et al. (2023)	Indonesia	University (teachers)	Qualitative	QuillBot, WordTune, Jenni, ChatGPT, Paperpal,	Teachers' perspective

No	Author	Country	Level	Design	Tool(s)	Focus
					Copy.ai, Essay Writer	
17	Mohammad et al. (2023)	Saudi Arabia	Undergraduate	Quasi-exp.	QuillBot	Writing outcomes (paraphrasing)
18	Kurniati & Fithriani (2022)	Indonesia	Postgraduate	Qualitative	QuillBot	Perception

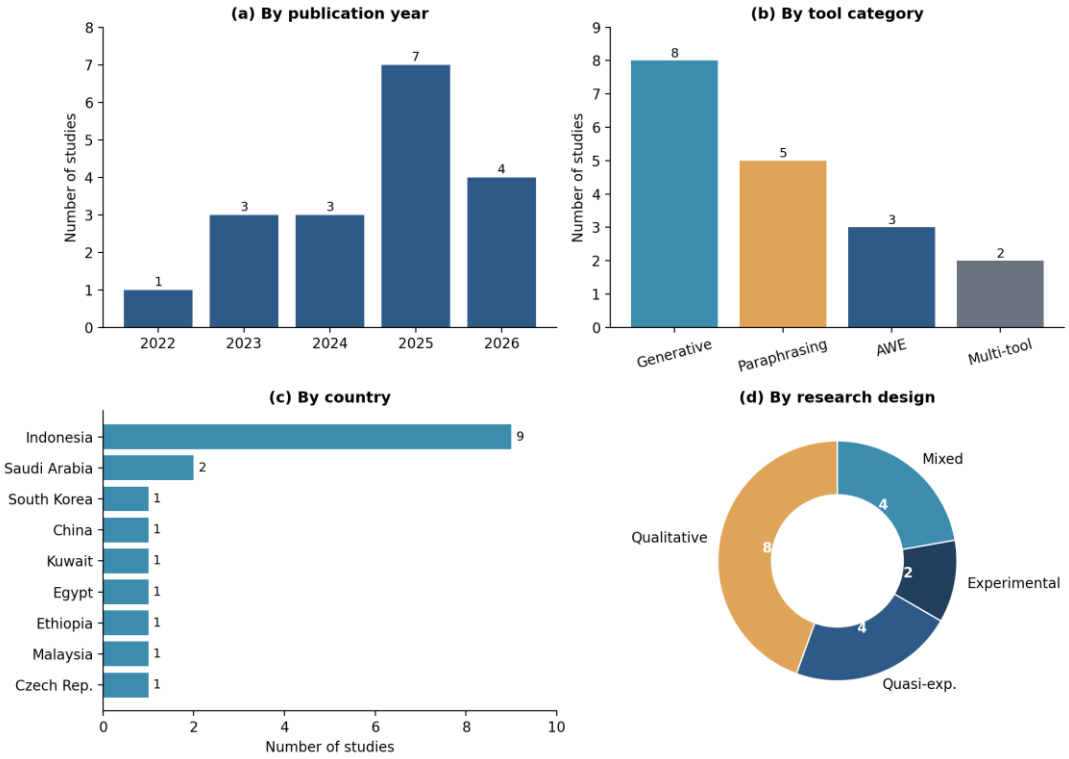

Figure 2. Distribution of the included studies by (a) publication year, (b) tool category, (c) country, and (d) research design (n = 18, current corpus).

As Figure 2 shows, generative-AI tools (n = 8) were the most frequently studied, followed by paraphrasing (n = 5), automated writing evaluation (n = 3), and multi-tool (n = 2) studies; research output rose sharply in 2025; Indonesia contributed the largest number of studies (n = 9); and perception-oriented qualitative and mixed-methods designs (n = 12) outnumbered studies with measured writing outcomes (n = 6). A concise, study-by-study summary of the findings across the three research questions is provided in Table 3 (at the end of this section).

3.2. RQ1: AI Tools Used and How They Are Used

Across the included studies, three broad categories of AI writing tools were evident, corresponding to the tool ecosystem outlined earlier.

Generative AI (ChatGPT and LLM-based tools). Generative tools featured in the largest number of studies and were used across multiple stages of the writing process — generating and developing ideas, drafting, revising, and obtaining on-demand feedback and grammatical correction (Polakova & Ivenz, 2024; Yang et al., 2025; Dzahabiyyah, 2025; Hayadi et al., 2025). Several studies emphasized more deliberate, strategic integration: Werdiningsih et al. (2024) described postgraduate students using ChatGPT alongside critical evaluation, peer consultation, and primary sources rather than as a wholesale substitute for their own writing. A distinct strand involved custom LLM-based systems purpose-built for the classroom, such as the scaffolding tool deployed by Myung et al. (2026).

Automated writing evaluation (AWE) and grammar-checking tools. AWE tools such as Grammarly (and comparable systems) were used primarily to provide automated corrective feedback on grammar, mechanics, and clarity, and to support cycles of revision (Wei et al., 2023; Mekheimer, 2025; Wale & Kassahun, 2024).

Paraphrasing tools. Paraphrasing tools, overwhelmingly QuillBot, were used chiefly at the revision and rewording stage — to reformulate sentences, improve word choice, and, in several studies, to help students avoid plagiarism (Nisa et al., 2026; Mohammad et al., 2023; Kurniati & Fithriani, 2022; Hamid et al., 2025). Notably, Nisa et al. (2026) characterized QuillBot as a "revision mediator," with students retaining authorship and critically vetting the tool's suggestions.

Finally, several studies documented multi-tool ecosystems rather than single tools. Marzuki et al. (2023), reporting from EFL teachers' perspectives, described the concurrent classroom use of QuillBot, WordTune, Jenni, ChatGPT, Paperpal, Copy.ai, and Essay Writer, while Alghasab (2025) reported students combining Grammarly, ChatGPT, and Google Translate. Across studies, tools were used both in teacher-mediated ways and through more autonomous, student-driven practices.

3.3. RQ2: Effects on Writing Quality and Skills

The reported effects of AI tools on writing revealed a central tension between improvements at the *surface* level and improvements in *higher-order* aspects of writing.

Surface-level gains. The most consistently reported benefits concerned grammatical accuracy, mechanics, and vocabulary. Studies of paraphrasing and grammar tools reported gains in synonym use, sentence structure, and word choice (Mohammad et al., 2023), and improvements in grammar and vocabulary were reported for generative tools as well (Dzahabiyyah, 2025). AWE studies similarly emphasized improvements in accuracy and error reduction (Wei et al., 2023).

Higher-order gains. A number of studies — frequently those with stronger designs or newer generative models — reported effects extending beyond the surface. Mekheimer (2025) reported gains in content, organization, and cohesion under a controlled design; Yang et al. (2025) found that the AI-supported group made more macro-level, content-oriented revisions; and Marzuki et al. (2023) reported unanimous agreement among teachers that AI tools enhanced the clarity and logical progression of students' writing — that is, aspects of content and organization. Werdiningsih et al. (2024) likewise associated strategic, critically mediated use with deeper engagement in the writing process.

The moderating role of design and evidence strength. The distinction between measured and perceived effects was important. The strongest evidence came from experimental and quasi-experimental studies, which generally found that groups using AI tools outperformed comparison groups on writing outcomes (Wei et al., 2023; Mekheimer, 2025; Wale & Kassahun, 2024; Yang et al., 2025; Mohammad et al., 2023). Perception studies were more numerous and almost uniformly positive, but rested on self-report and small samples, and therefore warranted more cautious interpretation. At least one quasi-experimental study illustrated methodological limitations that temper its conclusions, using the tool as both intervention and evaluator without a control group (Polakova & Ivenz, 2024).

Affective outcomes. Beyond text quality, several studies reported affective benefits, including increased writing self-efficacy (Yang et al., 2025), greater confidence (Dzahabiyyah, 2025), and more positive attitudes toward writing (Kurniati & Fithriani, 2022).

Uneven effects across learners. Importantly, the benefits were not uniform. The classroom deployment study by Myung et al. (2026) found that while LLM scaffolding improved grammatical accuracy overall, its step-by-step guidance demotivated lower-proficiency students and increased their reliance on the system, with lower-proficiency learners tending to outsource higher-order cognitive tasks while higher-proficiency learners delegated only lower-level ones. This suggests that the question is not only whether AI tools help, but *for whom* and *at what cost*.

3.4. RQ3: Challenges, Limitations, and Ethical Considerations

Alongside reported benefits, the included studies consistently surfaced a set of challenges and concerns.

Overreliance and reduced critical engagement. The most frequently raised concern was student overreliance. Teachers in Marzuki et al. (2023) worried that dependence on AI tools could inhibit students' critical thinking and problem-solving, and learner-focused studies echoed concerns about dependency (Hayadi et al., 2025). Myung et al. (2026) provided empirical, classroom-based evidence that scaffolding could foster dependency, particularly among weaker students.

Academic integrity. Several studies raised integrity issues, especially around paraphrasing tools. While QuillBot was frequently framed as a means of avoiding plagiarism, this framing sits in an ethical gray zone, since tool-driven rewording can shade into patchwriting rather than genuine reformulation. In contrast, studies such as Nisa et al. (2026) described a healthier model in which students maintained authorship and critically evaluated suggestions.

Accuracy and trust. Learners and teachers also raised concerns about the accuracy and reliability of AI outputs, including inaccurate suggestions (Hayadi et al., 2025), and ethical considerations around appropriate use were foregrounded in studies that examined them directly (Alsadoon, 2026; Rahayu et al., 2025).

Equity and classroom dynamics. A less commonly examined but significant set of concerns related to equity and social dynamics. Myung et al. (2026) documented that AI support could reshape classroom interaction — introverted students asked fewer questions, peer-learning interactions declined, and struggling students' difficulties were masked by AI-polished output, making them harder for teachers to identify.

Taken together, the included studies suggest that the effective use of AI tools in EFL/ESL writing depends heavily on how they are integrated pedagogically, with recurring calls for teacher mediation, explicit guidance on appropriate use, and attention to the differential impact on learners of varying proficiency.

Table 3. Summary of findings by research question (n = 18).

No	Author	RQ1 — Tool & use	RQ2 — Effects on writing	RQ3 — Challenges / ethics
1	Polakova & Ivenz (2024)	ChatGPT for feedback/practice	Reported writing improvement	Weak design (tool also grader; no control) limits claims
2	Rahayu et al. (2025)	ChatGPT acceptance (TAM)	Perceived usefulness; peer influence strong	Ethics/integrity concerns; proper stats
3	Yang et al. (2025)	ChatGPT for revision (10 wks)	More macro/content revisions; higher self-efficacy	Challenges addressed at surface level
4	Alsadoon (2026)	ChatGPT (students + teachers)	Perceived benefits for writing	Strong focus on ethics/appropriate use
5	Wale & Kassahun (2024)	Writerly (AWE) in drafting	AI group significantly outperformed control (t-test)	—
6	Mekheimer (2025)	Grammarly vs traditional (random assignment)	Gains in content, organization, cohesion	—
7	Alghasab (2025)	Multi-tool use; sociocultural lens	Perceived practices/benefits (K-12)	—
8	Wei et al.	AWE feedback (RCT)	Significant improvement	—

No	Author	RQ1 — Tool & use	RQ2 — Effects on writing	RQ3 — Challenges / ethics
	(2023)		in accuracy (strongest quant.)	
9	Hayadi et al. (2025)	ChatGPT perceptions	Favorable perceptions	Inaccuracy; dependency
10	Nisa et al. (2026)	QuillBot at revision ('revision mediator')	Supports reformulation; students keep authorship	Healthy anti-overreliance model; critical vetting
11	Werdiningsih et al. (2024)	ChatGPT strategic use (critical eval, peer consult)	Deeper engagement in writing process	Healthy critical use
12	Indria et al. (2026)	QuillBot for thesis/proposal paraphrasing	Perceived help	'Avoid plagiarism' framing = integrity gray zone
13	Hamid et al. (2025)	QuillBot use	Positive perceptions	—
14	Dzahabiyyah (2025)	ChatGPT use	Higher grammar/vocab/organization/confidence	—
15	Myung et al. (2026)	Custom LLM scaffolding, real-time (deployment)	Higher grammatical accuracy, BUT uneven (weak students)	STRONGEST: dependency, demotivation, equity/social costs
16	Marzuki et al. (2023)	Richest multi-tool inventory (teachers' view)	Improved clarity & logical progression (content/org.)	Overreliance -> may inhibit critical thinking
17	Mohammad et al. (2023)	QuillBot for paraphrasing (quasi-exp.)	Measured gains: synonyms, sentence structure, word choice	—
18	Kurniati & Fithriani (2022)	QuillBot in academic writing (postgrad)	Positive; 3 benefits (attitude, features, language dev.)	—

Note. "—" indicates that challenges were not prominent in that study (not that none exist). Studies with measured outcomes (Wei et al.; Mekheimer; Wale & Kassahun; Yang et al.; Mohammad; and the Myung deployment) carry greater evidentiary weight for effect claims than perception-only studies.

4. Discussion

4.1. Summary of Principal Findings

This review synthesized 18 primary empirical studies on the use of AI tools in EFL/ESL academic writing instruction. Three broad findings emerged. In response to RQ1, three categories of tools — generative AI, automated writing evaluation (AWE) and grammar-checking tools, and paraphrasing tools — were frequently used in combination and across multiple stages of the writing process, from idea generation and drafting to revision. In response to RQ2, the tools were consistently associated with gains in surface-level features such as grammar, mechanics, and vocabulary, whereas gains in higher-order dimensions such as content, organization, and coherence were reported chiefly by studies with stronger research designs or newer generative models. In response to RQ3, the studies repeatedly

raised concerns about overreliance, academic integrity, and the accuracy of AI output, along with a less frequently examined but consequential set of concerns about equity and classroom dynamics.

4.2. Beyond "Surface" and "Higher-Order": Which Tool, for Which Learner, at What Cost

The most salient pattern in the evidence is the tension between surface-level and higher-order effects on writing. This review suggests that the distribution of these effects is not random but appears to track two moderators. The first is tool type and model generation: grammar-focused AWE tools and paraphrasing tools were associated mainly with improvements at the surface — accuracy, mechanics, and word choice (e.g., Wei et al., 2023; Mohammad et al., 2023) — whereas newer generative models used under more controlled conditions were more likely to yield higher-order gains in content, organization, and cohesion (Mekheimer, 2025; Yang et al., 2025). The second moderator is research design: the smaller set of experimental and quasi-experimental studies with measured outcomes more often demonstrated genuine improvement, while the more numerous perception studies, although almost uniformly positive, provided weaker evidence for actual gains.

This pattern carries an important implication for interpreting the broader literature. Because the capabilities of these tools have changed rapidly, reported effects are to a considerable degree model-dependent, and conclusions drawn from earlier or less capable systems should not be assumed to hold for the tools currently in use. Reviews and primary studies alike therefore need to be read with attention to which specific tool and model version generated a given finding.

Most importantly, the classroom deployment evidence reframes the central question. Even where AI support improved grammatical accuracy, its benefits were unevenly distributed: step-by-step scaffolding demotivated lower-proficiency students and increased their reliance on the system, while their difficulties were masked from teachers by AI-polished output (Myung et al., 2026). The pertinent question therefore shifts from whether AI tools improve writing to a more differentiated one — which tool, used at which stage of the writing process, for which learner, and with what pedagogical framing, and at what cost to learner autonomy and classroom equity. Framed this way, the surface–higher-order debate becomes only one part of a larger question about the conditions under which AI tools genuinely serve EFL/ESL writing development.

4.3. Positioning Relative to Prior Reviews

The findings of this review both converge with and extend existing reviews of AI in academic writing. They converge in affirming that AI tools can meaningfully support EFL/ESL writing while simultaneously raising concerns about overreliance and academic integrity, echoing earlier syntheses (Alsaedi, 2024; Teng, 2024; Gabay et al., 2026). They extend prior work in four respects that correspond to the gaps identified at the outset.

- First, whereas tool-specific reviews examine a single tool — ChatGPT (Alsaedi, 2024; Teng, 2024) or Grammarly (Wijaya, 2025) — this review shows that such tools are used together in practice and that the surface–higher-order pattern differs by tool type, a distinction that single-tool reviews cannot reveal.
- Second, whereas broad higher-education reviews synthesize academic writing across disciplines (Shabir, 2025; Sanz-Tejeda et al., 2025; Gabay et al., 2026), the EFL/ESL focus adopted here — extending to secondary and K–12 contexts — surfaces language-learner-specific dynamics, such as the prominent use of paraphrasing tools for second-language rewording and the equity concerns that arise in mixed-proficiency EFL classrooms.
- Third, relative to the closest predecessor, which concentrates on generative AI in academic writing with an emphasis on English language teaching (Chanpradit, 2025), this review spans a wider tool set (AWE and paraphrasing tools as well as generative AI) and a longer time window (2019–2025, including the pre-ChatGPT AWE period).
- Fourth, by appraising the methodological quality of the included studies, this review makes explicit what many prior reviews leave implicit: a substantial share of the optimistic evidence rests on small, self-report perception studies, so conclusions about actual effects must be drawn cautiously. In this respect the synthesis is consistent with reviews finding that AI tools support structure and organization but are weaker on idea development, and are best positioned as a complement to — rather than a replacement for — teacher feedback (Cerón Urzúa et al., 2025).

4.4. Implications for EFL/ESL Writing Instruction

Several implications for practice follow from these findings.

- First, tool selection can be aligned with instructional goals: AWE and paraphrasing tools may be most useful for surface-level polishing and reformulation, whereas generative tools may better support idea development and higher-order revision — provided their use is carefully framed.
- Second, and consistently across studies, effective use depends on teacher mediation rather than unguided adoption; teachers play a central role in integrating tools purposefully and in helping students interpret and evaluate AI output (Marzuki et al., 2023; Werdiningsih et al., 2024).
- Third, explicit guidance on appropriate use and academic integrity is needed, particularly for paraphrasing tools, where reformulation can shade into patchwriting; instruction should cultivate genuine paraphrasing and critical vetting rather than uncritical acceptance of tool output (Nisa et al., 2026).
- Fourth, the uneven effects across learners imply that support should be differentiated by proficiency, since scaffolding that benefits stronger writers may foster dependency and demotivation among weaker ones (Myung et al.,

2026). Taken together, these implications point toward a model of strategic, critically mediated use — in which learners retain authorship and agency — as the most promising way to harness AI tools for writing development.

4.5. Limitations of This Review

This review has several limitations that should be considered when interpreting its findings.

- First, the evidence base is skewed toward perception and self-report studies, with a relatively small number of rigorous experimental designs; claims about the actual effects of AI tools on writing quality are therefore tentative and warrant confirmation through more controlled research.
- Second, the included studies were geographically concentrated, with a strong representation of research from Indonesia and the Global South; while this helps address a gap in the literature, it also limits the generalizability of the findings to other contexts.
- Third, the review was restricted to publications in English, introducing a potential language bias.
- Fourth, because the capabilities of AI tools are evolving rapidly, findings tied to particular tools or model versions may date quickly, and evidence from earlier systems may not transfer to current ones.
- Fifth, the methodological heterogeneity of the included studies precluded a meta-analysis, so the synthesis is necessarily narrative.

Finally, as with any systematic review, the conclusions are bounded by the search strategy and inclusion criteria adopted; studies outside the searched databases, the selected time window, or the language restriction were not captured.

5. Conclusion

This systematic literature review examined the use of AI tools in EFL/ESL academic writing instruction across studies published between 2019 and 2025. The evidence indicates that three categories of tools — automated writing evaluation and grammar-checking tools, paraphrasing tools, and generative AI — are increasingly used, often in combination and across all stages of the writing process. Their reported effects reveal a persistent tension: gains at the surface level of grammar, mechanics, and vocabulary are widely documented, whereas gains in higher-order dimensions such as content, organization, and coherence are reported mainly by studies employing stronger research designs or newer generative models. At the same time, the studies consistently raise concerns about overreliance, academic integrity, and, importantly, the uneven distribution of benefits, with scaffolding sometimes fostering dependency and demotivation among lower-proficiency learners. Beyond synthesizing these findings, the review contributes a broader, EFL-specific perspective that spans multiple tool types and educational levels, appraises the methodological quality of the evidence, and reframes the central question from whether AI tools help to which tool, for which learner, at which stage, and at what

cost. The findings suggest that the value of AI tools depends less on the technology itself than on how it is pedagogically mediated, pointing toward a model of strategic, teacher-guided, and critically evaluated use in which learners retain authorship and agency. Future research should prioritize rigorous, controlled studies across diverse contexts and proficiency levels, with explicit attention to equity and to the rapidly evolving capabilities of the tools themselves.

6. References

Alghasab, M. B. (2025). EFL secondary school students' use of AI tools for developing writing skills: Practices and perceptions. *Cogent Education, 12*(1), Article 2505304. https://doi.org/10.1080/2331186X.2025.2505304

Alsadoon, R. (2026). Perceptions vs. practices: Academic integrity and actual ChatGPT use among EFL students. *Frontiers in Psychology, 17*, Article 1796737. https://doi.org/10.3389/fpsyg.2026.1796737

Alsaedi, N. (2024). ChatGPT and EFL/ESL writing: A systematic review of advantages and challenges. *English Language Teaching, 17*(5), 41–50. https://doi.org/10.5539/elt.v17n5p41

Barrot, J. S. (2022). Integrating technology into ESL/EFL writing through Grammarly. *RELC Journal, 53*(3), 764–768. https://doi.org/10.1177/0033688220966632

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, Article 100745. https://doi.org/10.1016/j.asw.2023.100745

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cerón Urzúa, C. A., Ranjan, R., Méndez Saavedra, E. E., Badilla-Quintana, M. G., Lepe-Martínez, N., & Philominraj, A. (2025). Effects of AI-assisted feedback via generative chat on academic writing in higher education students: A systematic review of the literature. *Education Sciences, 15*(10), Article 1396. https://doi.org/10.3390/educsci15101396

Chanpradit, T. (2025). Generative artificial intelligence in academic writing in higher education: A systematic review. *Edelweiss Applied Science and Technology, 9*(4), 889–906. https://doi.org/10.55214/25768484.v9i4.6128

De Wilde, V. (2023). Lexical characteristics of young L2 English learners' narrative writing at the start of formal instruction. *Journal of Second Language Writing, 59*, Article 100960. https://doi.org/10.1016/j.jslw.2022.100960

Dzahabiyyah, L. (2025). Exploring the role of ChatGPT in enhancing EFL students' writing skills: A qualitative study at Universitas Negeri Semarang. *Journal of*

English Language and Education, 10(6), 868–880. https://doi.org/10.31004/jele.v10i6.1560

Gabay, R. A. E., Funa, A. A., & Ricafort, J. D. (2026). Generative artificial intelligence (GenAI) for academic writing in higher education: A scoping review of applications, challenges, and implications. *International Journal of Education in Mathematics, Science and Technology, 14*(1), 200–232. https://eric.ed.gov/?id=EJ1494380

Ghufron, M. A., & Rosyida, F. (2018). The role of Grammarly in assessing English as a foreign language (EFL) writing. *Lingua Cultura, 12*(4), 395–403. https://doi.org/10.21512/lc.v12i4.4582

Hamid, S. M., Danti, S., & Muhsin, M. A. (2025). EFL Indonesian students' voices: Quilbott as a viable an AI assistance in writing. *International Journal of Multicultural and Multireligious Understanding, 12*(3), 210–219. https://doi.org/10.18415/ijmmu.v12i3.6546

Hayadi, A., Sofyan, D., & Syafryadin. (2025). EFL students' perceptions toward ChatGPT as a writing instrument. *Eltin Journal: Journal of English Language Teaching in Indonesia, 13*(1), 75–88. https://e-journal.stkipsiliwangi.ac.id/index.php/eltin/article/view/5466

Hayes, J. R., & Flower, L. S. (2016). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Routledge.

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information, 34*(4), 285–291. https://doi.org/10.3233/EFI-180221

Hyland, K. (2019). *Second language writing* (2nd ed.). Cambridge University Press.

Indria, F., Hasanah, A., & Riznanda, W. A. (2026). Exploring EFL students' perceptions of QuillBot as an AI-powered paraphrasing tool in thesis writing. *English Education: Journal of English Teaching and Research, 11*(1). https://doi.org/10.29407/jetar.v11i1.27240

Kurniati, E. Y., & Fithriani, R. (2022). Post-graduate students' perceptions of QuillBot utilization in English academic writing class. *Journal of English Language Teaching and Linguistics, 7*(3), 437–451. https://doi.org/10.21462/jeltl.v7i3.852

Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL

teachers' perspective. *Cogent Education, 10*(2), Article 2236469. https://doi.org/10.1080/2331186X.2023.2236469

Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: Proficiency, revision frequency and writing quality. *Discover Education, 4*, Article 170. https://doi.org/10.1007/s44217-025-00602-7

Mohammad, T., Alzubi, A. A. F., Nazim, M., & Khan, S. I. (2023). EFL paraphrasing skills with QuillBot: Unveiling students' enthusiasm and insights. *Journal of Pedagogical Research, 7*(5), 359–373. https://doi.org/10.33902/JPR.202324645

Myung, J., Lim, H., Oh, H., Jin, H., Kang, N., Ahn, S.-Y., Hong, H., Oh, A., & Kim, J. (2026). When scaffolding breaks: Investigating student interaction with LLM-based writing support in real-time K-12 EFL classrooms. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (Article 838). Association for Computing Machinery. https://doi.org/10.1145/3772318.3791517

Nisa, V. Z., Riyono, A., & Syafei, M. (2026). Exploring EFL undergraduate students' experiences of using QuillBot AI in academic writing: An interpretive phenomenological study. *Edulitics Journal, 11*(1), 82–93. https://doi.org/10.52166/edulitics.v11i1.13216

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education, 11*(1), Article 2410101. https://doi.org/10.1080/2331186X.2024.2410101

Rahayu, A., Tarihoran, N., Rahmawati, E., Muslihah, E., Ma'mur, I., & Anita, A. (2025). Navigating the digital writing landscape: EFL students' perspectives on ChatGPT utilization. *International Journal of Learning, Teaching and Educational Research, 24*(5), 727–752. https://doi.org/10.26803/ijlter.24.5.37

Sanz-Tejeda, A., Domínguez-Oller, J. C., Baldaquí-Escandell, J. M., Gómez-Díaz, R., & García-Rodríguez, A. (2025). The impact of generative AI on academic reading and writing: A synthesis of recent evidence (2023–2025). *Frontiers in Education, 10*, Article 1711718. https://doi.org/10.3389/feduc.2025.1711718

Shabir, A. (2025). Generative AI writing tools and academic writing in higher education: A systematic review of empirical and review studies. *Information*

Technology Education Journal, 4(3), 543–557. https://doi.org/10.59562/intec.v4i3.10494

Sun, T., & Wang, C. (2020). College students' writing self-efficacy and writing self-regulated learning strategies in learning English as a foreign language. *System, 90*, Article 102221. https://doi.org/10.1016/j.system.2020.102221

Teng, M. F. (2024). A systematic review of ChatGPT for English as a foreign language writing: Opportunities, challenges, and recommendations. *International Journal of TESOL Studies, 6*(3), 36–57. https://doi.org/10.58304/ijts.20240304

United Nations. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development* (A/RES/70/1). https://sdgs.un.org/2030agenda

Vasylets, O., & Marín, J. (2022). Pen-and-paper versus computer-mediated writing modality as a new dimension of task complexity. *Languages, 7*(3), Article 195. https://doi.org/10.3390/languages7030195

Vu, P. T. (2024). Using Grammarly in enhancing students' grammatical accuracy in English writing: A case at a private college in Hanoi. *Proceedings of the AsiaCALL International Conference, 6*, 81–96. https://doi.org/10.54855/paic.2466

Wale, B. D., & Kassahun, Y. F. (2024). The transformative power of AI writing technologies: Enhancing EFL writing instruction through the integrative use of Writerly and Google Docs. *Human Behavior and Emerging Technologies, 2024*(1), Article 9221377. https://doi.org/10.1155/2024/9221377

Wei, P., Wang, X., & Dong, H. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. *Frontiers in Psychology, 14*, Article 1249991. https://doi.org/10.3389/fpsyg.2023.1249991

Werdiningsih, I., Marzuki, Indrawati, I., Rusdin, D., Ivone, F. M., Basthomi, Y., & Zulfahreza. (2024). Revolutionizing EFL writing: Unveiling the strategic use of ChatGPT by Indonesian master's students. *Cogent Education, 11*(1), Article 2399431. https://doi.org/10.1080/2331186X.2024.2399431

Wijaya, K. F. (2025). The role of Grammarly as an automated writing evaluation tool in EFL writing learning process. *Datokarama English Education Journal, 6*(2), 61–76. https://doi.org/10.24239/dee.v6i2.127

Yang, Y., Hussin, S., & Hashim, H. (2025). Revising with intelligence: ChatGPT feedback and its impact on EFL students' revision and self-efficacy. *Forum for Linguistic Studies, 7*(7), 352–367. https://doi.org/10.30564/fls.v7i7.9845