Vol.11, No. 1, 2026, pp. 899-913
DOI: https://doi.org/10.29210/30036949000

Contents lists available at Journal ELORA

JRTI (Jurnal Riset Tindakan Indonesia)
ISSN: 2502-079X (Print) ISSN: 2503-1619 (Electronic)

Journal homepage: https://jrti.eloracenter.org/jrti

Large language models for english learning and writing: a systematic review for sport sciences

Kristian Burhan*), M. Arie Desman
Universitas Negeri Padang, Indonesia

Article Info

Article history:

Received Aug 22th, 2026
Revised Aug 25th, 2026
Accepted Aug 27th, 2026

Keyword:

Generative artificial intelligence,
Large language models,
English for specific purposes,
Language assessment,
Academic writing

ABSTRACT

This systematic review synthesized empirical evidence on large language models (LLMs) and generative AI for English-language learning, assessment, and academic writing, with implications for English for specific purposes (ESP) in sport sciences. Following PRISMA 2020, a structured Scopus TITLE-ABS-KEY search identified 871 records; 50 peer-reviewed journal studies published between 2024 and 2026 met the eligibility criteria. Studies were synthesized using a three-stage thematic approach based on coding, descriptive themes, and analytical themes. Methodological reporting and applicability were appraised using a transparent four-domain FICO rubric. Screening was conducted by one reviewer; therefore, inter-rater reliability statistics were not calculable retrospectively. Four analytical clusters were identified: language-skills learning (8 studies), assessment and evaluation (8), academic writing and feedback (30), and adoption, perception, and learner agency (4). Evidence generally favoured structured, teacher-mediated GenAI use, while LLM assessment showed high reliability in some rubric-based tasks but variable validity. GenAI was most consistently useful for feedback, revision, and self-regulated learning. Because designs and outcomes were heterogeneous, no pooled effect size was calculated. Future research should prioritise discipline-specific ESP studies, longitudinal designs, transparent validation, and responsible AI integration.

Corresponding Author:

Kristian Burhan,
Universitas Negeri Padang,
Email: kristianburhan@unp.ac.id

Introduction

English sits at the centre of scientific communication worldwide; it is, for all practical purposes, the lingua franca of the sciences. Command of academic and professional English increasingly determines whether researchers can publish, move between institutions, and take part in disciplinary life. In applied fields such as sport sciences, kinesiology, and health, the pressure is especially visible. Researchers and practitioners have to read, write, and present specialised English to disseminate evidence and engage international communities. Then generative artificial intelligence (GenAI) and large language models (LLMs) arrived, bringing tools that produce fluent, context-sensitive language on demand. Their arrival prompted a fundamental rethink of how English is taught, learned, and evaluated (Jeon, 2025; Kohnke et al., 2025). This is a macro-level shift. GenAI is no longer a marginal aid; it has become pervasive infrastructure reshaping literacy practices in higher education.

What interests this review is a narrower phenomenon: the pedagogical integration of GenAI into English-language learning, assessment, and academic writing. Since conversational LLMs were released publicly, educators have experimented rapidly with tools such as ChatGPT, Gemini, and Copilot, applying them to vocabulary acquisition, speaking practice, and writing feedback (Alpar, 2025; Tram et al., 2024). Enthusiasm,

though, has outrun the evidence. The field still lacks a coherent synthesis that separates robust findings from preliminary claims across the distinct functions these tools serve.

A growing body of empirical research now covers GenAI across the four language skills and the writing process. Studies report gains in vocabulary knowledge and retention (Abdelhalim & Alsehibany, 2025), in speaking proficiency (Behforouz et al., 2025; Kim, 2025), and in writing accuracy and fluency (Alasal, 2026; Almutairi et al., 2026). Qualitative work complements these results, documenting how learners interact with chatbots during composition (Huang & Wang, 2025). Reviews of recent scholarship confirm that the research base is accelerating, though they also warn about methodological heterogeneity (Jeon, 2025).

Methodological and technological advances have also expanded what GenAI can do in language education. Systems now combine LLMs with speech recognition and automated scoring to offer interactive speaking practice (Lu et al., 2026). Customised chatbots are being deployed for automated essay scoring (Lan et al., 2025), and multiple tools are being knitted into automated writing evaluation workflows (Gozali et al., 2024). Comparative evaluations of GPT-family and competing models have begun to quantify reliability and validity for assessment tasks (Pack et al., 2024). The field appears to be shifting from novelty demonstrations toward rigorous measurement.

Despite this momentum, an important gap concerns disciplinary specificity. Most evidence comes from general English as a foreign language (EFL) settings, whereas English for specific purposes (ESP) evidence for applied scientific fields remains comparatively thin. In this review, the sport-sciences construct is operationalised as discipline-specific English needs relevant to sport, kinesiology, exercise, and health-science study and professional communication, including domain vocabulary, professional speaking, English academic writing, revision, and formative assessment. The included evidence was not restricted to sport-science students because doing so would have excluded the broader EFL evidence needed to establish transferable pedagogical patterns. Accordingly, sport sciences is treated as the intended ESP translation context rather than as a mandatory population criterion. This distinction is important when interpreting the scope of the findings.

A second gap is theoretical and methodological. Many studies are short, single-group, or perception-based. The field has yet to converge on validated constructs for learner agency, feedback literacy, and self-regulated learning in GenAI-mediated environments (Hamamra et al., 2026). Evidence on how AI and teacher feedback should be sequenced or combined is emerging but fragmented (Maleki, 2026; Tran, 2025). Comparisons of human and AI assessment vary in rigour (Ennadir & Habiballah, 2025). These limitations complicate cumulative theory-building.

The pace of adoption makes a systematic synthesis urgent, all the more so given the stakes around academic integrity and equity. Learners are increasingly offloading writing tasks to GenAI (Lai et al., 2026), while institutions weigh trust in AI-generated output against concerns about over-reliance (Anwer, 2026). Decision-makers need consolidated, transparent evidence. This review responds to that need by mapping the field through the PRISMA 2020 framework and interpreting the findings for ESP contexts such as sport sciences English.

The synthesis is organised around three research questions. The first concerns learning outcomes across language skills, the second the reliability and validity of GenAI for assessment, and the third the role of GenAI in academic writing and feedback processes. Taken together, these three questions define the analytical scope and organise the thematic findings reported below.

RQ1: How does the use of generative AI and large language models affect English-language learning outcomes across vocabulary, speaking, listening, and reading skills?

RQ2: How reliable and valid are large language models when used for the assessment and automated evaluation of English-language performance?

RQ3: What roles do generative AI tools play in supporting English academic writing, feedback, and revision, and with what implications for learner agency and integrity?

To distinguish more robust findings from preliminary claims, the review prioritised empirical evidence with an identifiable intervention, exposure, or assessment application; an analysable language, assessment, or writing outcome; sufficient methodological description to interpret the design and outcome; and a peer-reviewed journal publication. Study size was not used as a universal exclusion threshold because qualitative, psychometric, and intervention studies address different questions. Instead, sample size, design strength, outcome validity, and replication potential were considered during FICO appraisal and narrative interpretation. Claims were labelled as more convincing when they were supported by controlled or comparative designs, validated instruments, repeated or delayed measures, or convergent evidence across studies; findings based primarily on perception, single-group designs, or very small samples were treated as preliminary.

The review answers these questions through a transparent and reproducible synthesis of 50 recent studies. In doing so, it makes three contributions: an integrated evidence map spanning learning, assessment, and writing; a critical appraisal of methodological maturity; and a research agenda that foregrounds discipline-specific applications, including English for sport sciences. What makes this review novel is the bridge it builds between general EFL evidence and the ESP requirements of applied scientific fields.

Method

This study is a systematic literature review designed to identify, appraise, and synthesize empirical evidence on LLM-supported English learning, assessment, and academic writing. Reporting followed PRISMA 2020. The protocol was not prospectively registered, and this is acknowledged as a limitation. The unit of inclusion was the peer-reviewed journal study, while review articles were used only for contextual comparison and were not included in the 50-study synthesis.

This study is built around a systematic literature review (SLR). We chose that design because the research questions call for a synthesis that is comprehensive, transparent, and reproducible across a rapidly expanding evidence base, rather than a primary empirical investigation. The SLR approach gives us a structured way to identify, appraise, and synthesize relevant studies, and the explicit protocols help keep selection bias in check. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement. That framework structures the stages of identification, screening, eligibility, and inclusion, and it supports methodological transparency throughout the review workflow.

For the search, we ran a structured Boolean query against the title, abstract, and keyword fields (TITLE-ABS-KEY) in Scopus. The query brought together three conceptual blocks: GenAI/LLM technologies, English-language learning, and learning, assessment, or writing outcomes. The final search syntax was preserved exactly as executed in the Scopus export. Truncation was used to capture morphological variants, and quotation marks preserved multi-word concepts. No citation chaining or supplementary database searching was used; this single-database design is a stated source of selection bias.

For the search, we ran a structured Boolean query against the title, abstract, and keyword fields (TITLE-ABS-KEY) in Scopus. The query brought together three conceptual blocks: generative AI/LLM technologies, the English language, and learning, assessment, or writing outcomes. Truncation was used so morphological variants would also be captured.

("generative AI" OR "large language model" OR "ChatGPT" OR "GPT-4" OR "Gemini" OR "Copilot") AND ("English" OR "EFL" OR "ESL" OR "ESP" OR "second language" OR "foreign language") AND ("writing" OR "assessment" OR "feedback" OR "vocabulary" OR "speaking" OR "learning")*

Truncation with the asterisk caught plural and derivational forms. Quotation marks, meanwhile, locked the search onto exact phrases. The TITLE-ABS-KEY field code did the rest, filtering matches down to substantive bibliographic fields. No additional citation-chaining databases were used, so every record traces back to a single, auditable source consistent with the export analysed here.

Scopus was the sole bibliographic source. It was selected because its multidisciplinary coverage spans applied linguistics, educational technology, computer-assisted language learning, and related educational research, while its structured export preserved bibliographic metadata needed for screening and descriptive analysis. The retrieval produced 871 records. No forward/backward citation chaining, grey-literature searching, or supplementary database search was undertaken. The exact search execution date was not retained in the source file and should be added by the authors in the journal submission record if available.

Scopus was the sole bibliographic source, and the primary one. Its coverage of peer-reviewed journals in applied linguistics, educational technology, and computer-assisted language learning is broad, and the structured metadata made reproducible screening a straightforward matter. The search was executed and exported as a comma-separated metadata file, which now forms the authoritative evidence base for this review. Authors, titles, journals, years, digital object identifiers: all of it came directly from that export.

Predefined inclusion and exclusion criteria governed eligibility (Table 1). To qualify, a study had to be an empirical, peer-reviewed journal article written in English, published between 2024 and 2026, directly examine a GenAI or LLM application in English-language learning, assessment, or writing, and report an outcome that could be interpreted from the full text. Conference papers, book chapters, editorials, protocols, reviews, and purely technical NLP papers without an English-education outcome were excluded from the synthesis.

Table 1. Inclusion and exclusion criteria.

Criterion	Inclusion	Exclusion
Language	English only	Non-English publications
Document type	Peer-reviewed journal article	Conference paper, book chapter, editorial, note
Publication period	2024–2026	Before 2024
Subject area	English-language education, ESP, CALL, educational technology	Unrelated disciplines (e.g., purely clinical or engineering NLP)
Technology focus	Identifiable GenAI / LLM tool or method	No GenAI or LLM focus
Outcome	Analysable empirical learning, assessment, or writing outcome	No empirical or analysable outcome

Selection happened in three sequential stages. Titles and abstracts were screened first; 775 of 871 records were excluded. Ninety-six reports were then retrieved and assessed in full, of which 46 were excluded, leaving 50 studies. Each exclusion was assigned a primary reason and documented in the screening record. Screening was conducted by one reviewer using the predefined criteria. A second independent screening record was not available, so Cohen's kappa and percentage agreement could not be calculated retrospectively; this limitation is stated explicitly rather than estimated. Ambiguous records were carried forward to full-text assessment rather than being excluded at title-abstract stage.

The FICO framework was used as a transparent appraisal rubric across the heterogeneous evidence base. Focus assessed whether the research question and psychological/educational construct were explicit; Information assessed reporting of sample, intervention, comparator, measures, and analytic procedures; Context assessed relevance of the educational and disciplinary setting; and Outcome assessed whether the reported findings were interpretable and directly connected to the stated outcome. Each domain was considered on a three-level qualitative scale (weak, adequate, strong) for reporting purposes. FICO was used as an applicability and reporting-quality framework rather than a validated risk-of-bias instrument, and no pooled quality score was used to weight effect sizes. Because the included designs ranged from qualitative case studies to quasi-experiments and psychometric comparisons, a single established risk-of-bias instrument was not imposed across all designs. This limitation is addressed in the Discussion.

A standardised extraction template was used for each included study. Fields covered author and year, study context and population, design, sample size, GenAI/LLM system, comparator where applicable, language skill or assessment target, intervention duration, outcome instrument, main quantitative or qualitative finding, and key limitations. For quantitative studies, the extraction recorded the reported statistic (e.g., mean difference, reliability coefficient, correlation, or effect-size estimate) when available. Because the source set did not provide a sufficiently complete and commensurate set of numerical estimates across all 50 studies, these values were not recomputed from incomplete information and no pooled effect size was generated. The extraction matrix supports the descriptive and thematic synthesis.

Network and Bibliometric Analysis Methodology

Alongside the qualitative synthesis, descriptive bibliometric analyses characterised the included corpus. Author-keyword co-occurrence frequencies were computed to surface the dominant conceptual clusters, and temporal and geographic distributions were tabulated to show how the field matured and how far its reach extended. These analyses were descriptive by design. They placed the thematic findings inside the broader structure of the literature, and they didn't test any bibliometric hypotheses.

Data Analysis and Synthesis

Findings from studies with different designs were integrated through a three-stage thematic synthesis. First, study findings were coded line by line around the review constructs. Second, codes were organised into descriptive categories. Third, those categories were mapped onto analytical themes aligned with RQ1-RQ3. Thematic counts were used descriptively, not as estimates of effect. No meta-analysis, I-squared calculation, pooled effect size, or publication-bias statistic was performed because the included studies differed in intervention type, outcome construct, measurement scale, comparator, and timing. The absence of statistical pooling is therefore a methodological decision based on non-comparable estimands, not an omission of analysis. Disconfirming findings were actively retained during synthesis.

Reporting and Documentation

The review follows the PRISMA 2020 reporting checklist across identification, screening, eligibility, and inclusion. A PRISMA 2020 flow diagram (Figure 1) tracks the flow of records. Every numerical value reported in the abstract, methods, and results matches the others, and each one derives from the analysed Scopus export.

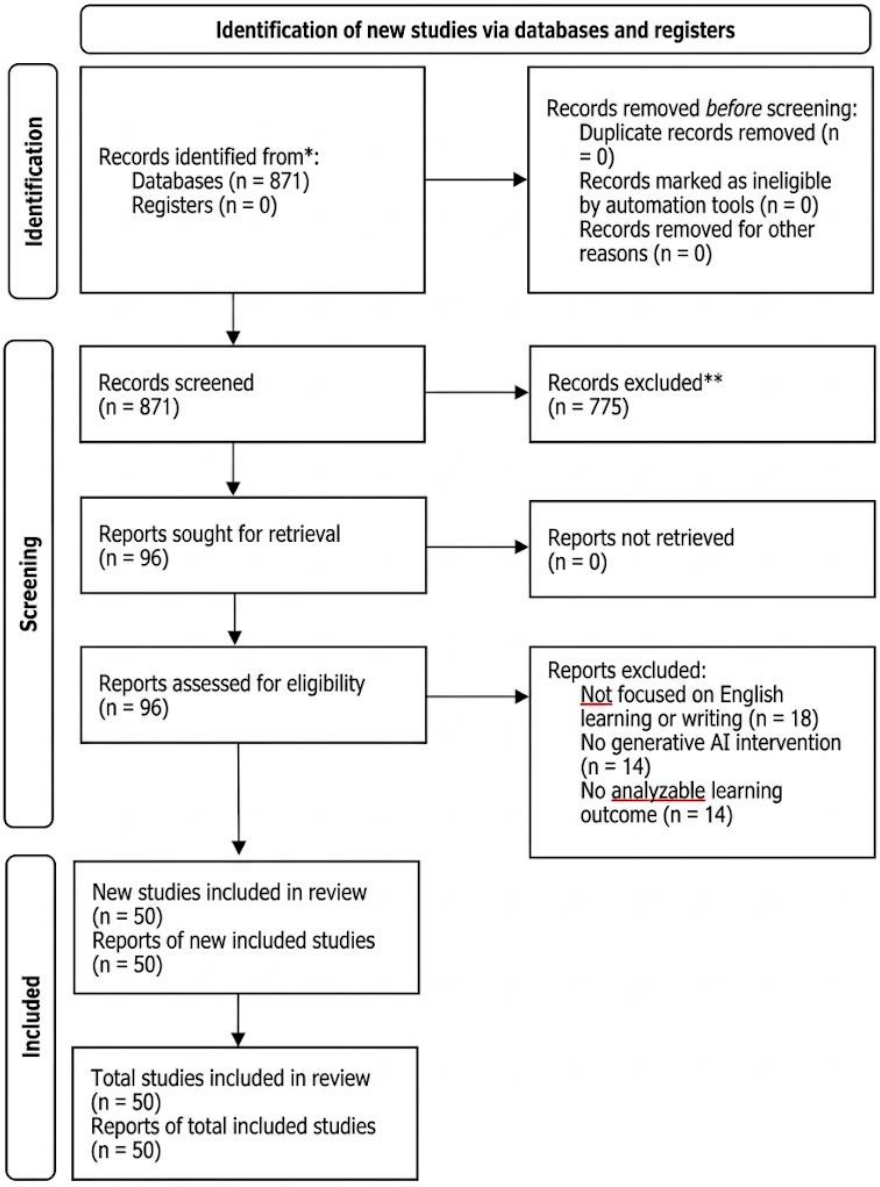

Figure 1. PRISMA 2020 flow diagram of the study identification, screening, eligibility, and inclusion process.

Results and Discussions

Study Selection Results

The Scopus search returned 871 records. No duplicate records were removed from the final export. Title and abstract screening excluded 775 records: 452 did not focus on English-language education, 233 lacked an identifiable GenAI or LLM focus, and 90 were outside the review scope. Ninety-six reports were retrieved and assessed in full. Full-text assessment excluded 46 reports: 18 were not centred on English learning or writing, 14 contained no GenAI intervention or analysis, and 14 lacked an analysable empirical outcome. Fifty studies met all eligibility criteria and were included in the thematic synthesis. These counts correspond exactly with Figure 1 and the study-selection narrative in Methods.

Descriptive Characteristics

The 50 included studies were published between 2024 and 2026, with publication activity concentrated in 2025 and 2026. Figure 2 shows 6 included studies from 2024, 24 from 2025, and 20 from 2026. Across thematic focus, 30 studies addressed academic writing and feedback, 8 addressed language-skills learning, 8 addressed assessment and evaluation, and 4 focused primarily on adoption, perception, or learner agency. This frequency pattern indicates that writing and feedback are currently the most empirically developed part of the literature, whereas listening, reading, and discipline-specific ESP applications remain less represented. Table 2 is retained

as a representative evidence table rather than an exhaustive inventory; the main analytical synthesis is reported below and individual-study details are secondary to the thematic results.

Table 2. Summary of included studies (representative sample, n = 10).

Title	Author(s)	Year	Country	Method	Key findings
Integrating ChatGPT for vocabulary learning and retention: A classroom-based study of Saudi EFL learners	Abdelhalim and Alsehibany	2025	Saudi Arabia	Quasi-experimental pre-/post-test with delayed post-test; explanatory sequential mixed methods; 71 EFL learners	ChatGPT integration significantly improved vocabulary knowledge and retention over teacher-led instruction; learners valued real-time, tailored feedback.
Exploring the efficacy of ChatGPT-4 feedback in second language Spanish writing	Barrios-Beltran	2025	Spain	Qualitative; thematic analysis of questionnaires, writing revisions, and learner reflections; advanced L2 Spanish learners	ChatGPT-4 feedback supported targeted revision, metacognitive reflection, and learner agency, complementing rather than replacing instructor input.
ChatGPT as an automated writing evaluation (AWE) tool: feedback literacy development and AWE tools' integration framework	Gozali et al.	2024	Indonesia	Qualitative single case study; interviews, reflective journals, e-portfolios; 18 EFL students	ChatGPT with Grammarly and Quillbot advanced learners' feedback literacy, with ChatGPT strongest in feedback processing; proposed an AWE integration framework.
Can ChatGPT serve as a writing collaborator? Insights from Chinese EFL learners	Huang and Wang	2025	China	Qualitative; interaction logs, think-aloud protocols, texts, interviews; 9 upper-intermediate learners; argumentative task	Student–ChatGPT collaboration shaped final text quality; learners positioned ChatGPT as a collaborator yet exercised selective uptake of suggestions.
Development of Linguistic Competence in English for Specific Purposes Through ChatGPT: A Case Study	Jordán Enamorado	2025	Not specified	Case study; ESP (English for Tourism); 91 university students; reading, vocabulary, and role-play tasks	ChatGPT enhanced domain-specific linguistic competence in an ESP setting, supporting authentic professional-language tasks and learner engagement.
An Intelligent English-Speaking Training System Using Generative AI and Speech Recognition	Lu et al.	2026	Taiwan	System design and evaluation; GenAI dialogue agents, speech recognition, automated scoring; non-native speakers	The GenAI speaking system improved oral proficiency through interactive practice, real-time scoring, and personalized feedback in a low-anxiety environment.
Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability	Pack et al.	2024	Not specified	Comparative reliability/validity study; four LLMs vs. two human raters; 119 placement essays scored twice	GPT-4 achieved excellent intra-rater reliability and good validity for automated essay scoring; models varied, and human oversight remained necessary.
Sustainable L2 writing pedagogy in Turkish higher education: Effects of AI-mediated feedback on self-regulated learning and writing performance	Selvi et al.	2026	Türkiye	Quasi-experimental, three feedback groups (Grammarly, ChatGPT, instructor); mixed-factorial ANOVA; 84 students; 10 weeks	AI-mediated feedback improved self-regulated learning and writing performance; GenAI feedback rivalled instructor feedback on several measures.
Enhancing EFL Writing Revision Practices: The Impact of AI- and Teacher-	Tran	2025	Vietnam	Preliminary experimental study of feedback sequences (teacher-first vs.	Feedback ordering shaped revision behaviour; combined teacher and AI feedback enhanced

Title	Author(s)	Year	Country	Method	Key findings
Generated Feedback and Their Sequences				AI-first); 14 undergraduates; 15 weeks	revision practices beyond either source alone.
Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments	Yavuz et al.	2025	Not specified	Reliability/validity study; ChatGPT and Bard vs. 15 EFL instructors; rubric with five domains; three essays	Fine-tuned ChatGPT reached very high reliability (ICC = 0.97); LLM and human grades overlapped substantially in several rubric domains.

Table 3. Classification of included studies by research design, theme, technology, and outcome.

Author(s)	Year	Country	Research design	Theme / focus	Technology	Outcome
Abdelhalim and Alsehibany	2025	Saudi Arabia	Quasi-experiment (mixed methods)	Vocabulary learning & retention	ChatGPT (tutoring & exercises)	Gains in vocabulary knowledge and delayed retention
Barrios-Beltran	2025	Spain	Qualitative case analysis	Written corrective feedback	ChatGPT-4	Targeted revision, metacognition, learner agency
Gozali et al.	2024	Indonesia	Qualitative case study	Automated writing evaluation	ChatGPT + Grammarly + Quillbot	Feedback-literacy development; integration framework
Huang and Wang	2025	China	Qualitative interaction study	Human–AI writing collaboration	ChatGPT	Collaboration patterns influencing text quality
Jordán Enamorado	2025	Not specified	Case study (ESP)	ESP / domain-specific competence	ChatGPT	Enhanced tourism-English competence
Lu et al.	2026	Taiwan	System design & evaluation	Speaking / oral proficiency	GenAI + speech recognition + AES	Improved oral proficiency; reduced anxiety
Pack et al.	2024	Not specified	Comparative psychometric study	Automated essay scoring	PaLM 2, Claude 2, GPT-3.5, GPT-4	GPT-4 best; reliability–validity trade-offs
Selvi et al.	2026	Türkiye	Quasi-experiment (ANOVA)	Feedback & self-regulated learning	ChatGPT vs. Grammarly vs. instructor	SRL and writing-performance gains
Tran	2025	Vietnam	Experimental (feedback sequencing)	Writing revision practices	Gemini + teacher feedback	Sequence effects on revision quality
Yavuz et al.	2025	Not specified	Psychometric reliability study	Rubric-based grading	ChatGPT & Bard	High grading reliability vs. human raters

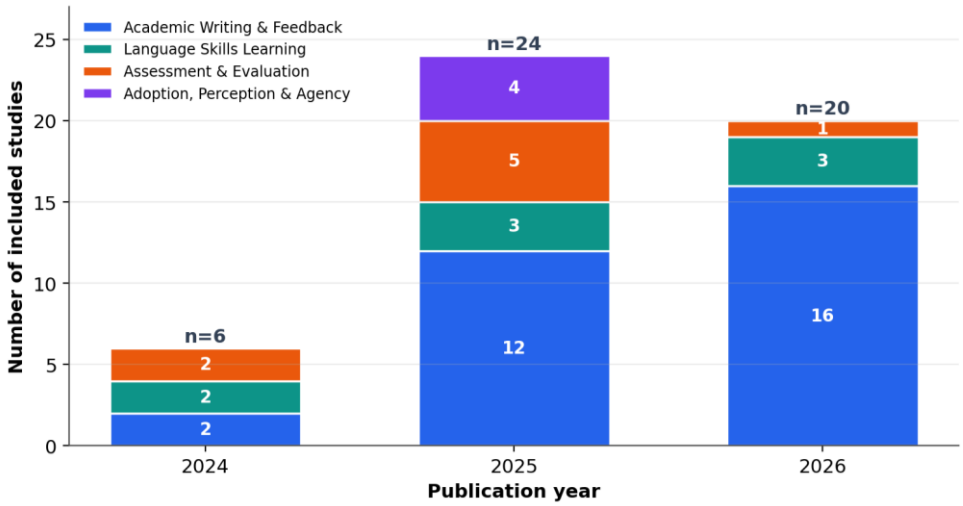

Figure 2. Temporal distribution of included studies by thematic focus (2024–2026).

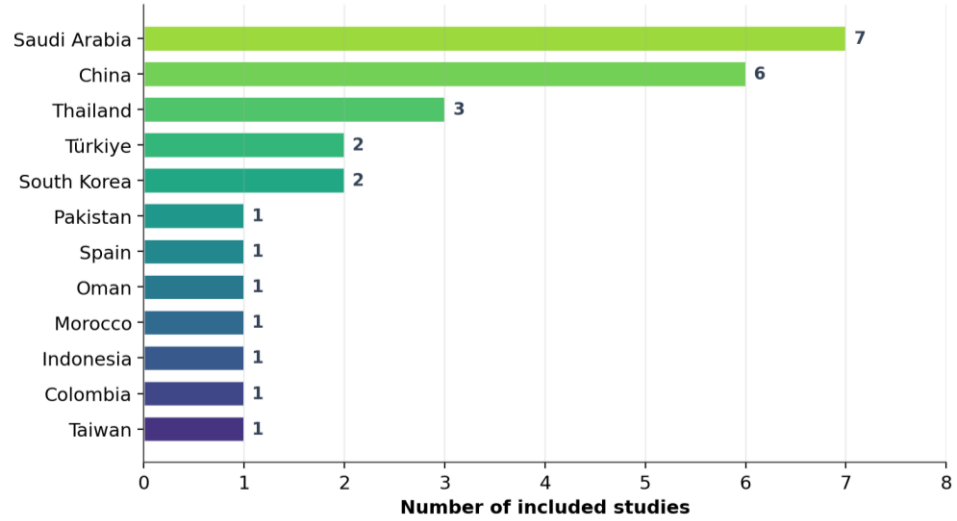

Figure 3. Geographic distribution of included studies by country of study context.

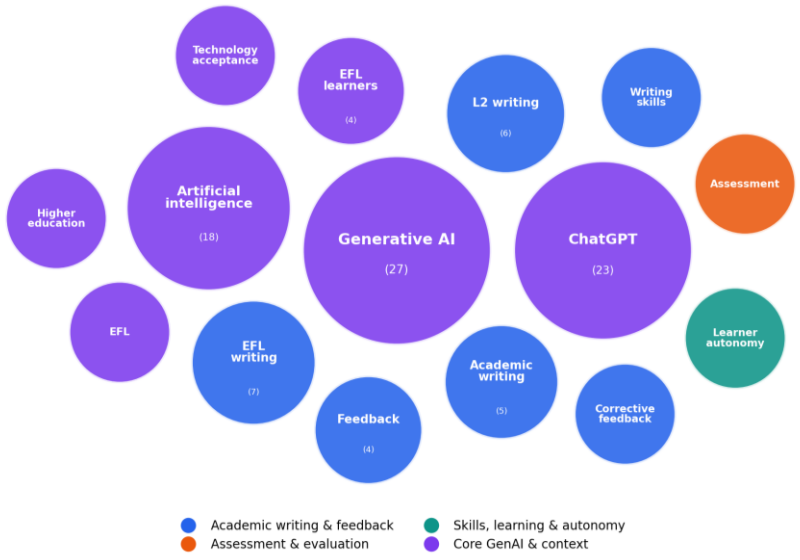

Figure 4. Author-keyword co-occurrence map across four thematic clusters (bubble size denotes frequency).

Thematic Synthesis

Findings for RQ1: Learning Outcomes Across Language Skills

Across the corpus, measurable gains in discrete language skills were consistently linked to GenAI tools, with vocabulary and productive skills showing the strongest patterns. Abdelhalim and Alsehibany (2025), working from a quasi-experimental classroom design, reported that ChatGPT integration significantly improved vocabulary knowledge and delayed retention when compared with teacher-led instruction. They attributed that effect to real-time, individualised feedback and adaptive exercises. Elsewhere, learner perceptions of ChatGPT for vocabulary work came out positive (Almarwani & Alhenaky, 2026), and spelling-focused interventions suggest LLMs can cut specific error types among EFL learners (Jaashan & Alashabi, 2025).

Oral skills showed a similar pattern. When LLMs were combined with speech recognition and automated scoring, interactive speaking practice could take place in low-anxiety environments (Lu et al., 2026), and mobile AI tutoring built on ChatGPT lifted speaking performance (Behforouz et al., 2025). Gains in speaking confidence and engagement were reported for elementary and general learners as well (Tai & Chen, 2024), while broader learning-support designs improved listening and speaking outcomes (Kim, 2025). Outside the classroom, self-directed use of ChatGPT allowed autonomous English practice (Tram et al., 2024).

Across the 50-study corpus, the strongest recurring outcome pattern was improvement in productive English performance and feedback-supported revision, while durable gains in receptive skills were less frequently tested. The thematic distribution is therefore not simply a count of positive studies: it reflects the concentration of the literature on writing and feedback. This distinction matters because writing outcomes have been measured more often and with more varied rubrics than listening or reading outcomes. Claims of broad language proficiency improvement should therefore be interpreted as domain-specific rather than universal.

Still, the picture is more nuanced. Gains tended to be strongest when GenAI was embedded in structured, teacher-designed tasks; used in isolation, it delivered far less. Several studies also emphasised learner perceptions and engagement over durable proficiency gains (Zheldibayeva, 2025). Reading-focused work, including ChatGPT-mediated mind mapping, looked promising for comprehension but rested on small samples (Alyami, 2026). Personalised teaching applications point toward differentiation potential (Kohnke et al., 2025), though the heterogeneity of designs makes broad generalisation risky. On balance, RQ1 gets an affirmative but conditional answer: GenAI supports skill development, particularly with pedagogical scaffolding in place.

Findings for RQ2: Reliability and Validity of GenAI Assessment

A good number of studies turned GenAI toward assessment, and the results were encouraging, if qualified. Yavuz et al. (2025), using rubric-based grading, reported very high reliability for a fine-tuned ChatGPT model, with an intraclass correlation of 0.97 and substantial agreement between LLM and human raters across several domains. Pack et al. (2024) similarly found that GPT-4 showed excellent intra-rater reliability and good validity when scoring essays written by English-language learners, outperforming earlier models.

Beyond the coefficients, several studies looked at how automated evaluation fits into instruction. Distance settings saw customised GenAI chatbots deployed for automated essay scoring (Lan et al., 2025), and comparative work weighed human and AI-assisted assessment to see where they converge and diverge (Ennadir & Habiballah, 2025). These findings were extended to authentic institutional settings through automated evaluation of ESL writing in English-medium instruction (Shahzad & Abbas, 2026). Meanwhile, automated writing evaluation frameworks placed ChatGPT alongside established tools in order to build learner feedback literacy (Gozali et al., 2024).

The evidence on assessment should be separated into reliability and validity. Reliability was often high in rubric-based scoring tasks: for example, Yavuz et al. (2025) reported ICC = 0.97, while Pack et al. (2024) reported strong reliability for GPT-4. These coefficients describe agreement or consistency within the tested scoring conditions; they do not establish validity for every learner population, rubric, genre, or high-stakes decision. Accordingly, the synthesis treats reliability as an encouraging but insufficient condition for automated assessment.

Validity, though, stayed model- and task-dependent, and human oversight was almost always recommended. Reliability shifted across models and rubric domains. Even with high surface accuracy, LLMs may misjudge higher-order qualities like argumentation and coherence (Pack et al., 2024). For RQ2, then, the synthesis offers measured optimism: contemporary LLMs, especially fine-tuned or GPT-4-class models, can reach high reliability in rubric-based scoring, yet valid use calls for human calibration, transparent rubrics, and care in high-stakes settings.

Findings for RQ3: GenAI in Academic Writing, Feedback, and Revision

The largest thematic cluster was academic writing and feedback. Following AI-assisted instruction, multiple studies documented better writing accuracy and fluency (Almutairi et al., 2026; Franjić & Božinovski, 2025). Controlled comparisons went further, showing that GenAI feedback improved writing performance and self-regulated learning, at times rivalling instructor feedback (Selvi et al., 2026). ChatGPT-generated feedback, specifically, produced gains in writing skills and noticeable shifts in learner perception (Alsofyani & Barzanji, 2025), and preliminary evidence on proficiency and revision frequency reinforced the pattern (Mekheimer, 2025).

A recurring idea was to frame GenAI as a collaborator rather than an author. Huang and Wang (2025) found that Chinese EFL learners used ChatGPT as a writing collaborator while selectively taking up its suggestions. Qualitative work on AI-generated feedback made a similar point, emphasising its role in metacognition and learner agency (Barrios-Beltran, 2025). Feedback-focused studies asked how AI and teacher feedback could be combined and sequenced to make revision more effective (Maleki, 2026; Tran, 2025), how teacher-moderated GenAI feedback shapes revision quality (Pretorius & Thewissen, 2026), and which tasks learners prefer to offload to GenAI (Lai et al., 2026). Chatbots further acted as revision aids, pulling attention away from surface error correction and toward higher-order concerns (Chon & Shin, 2026).

Across the writing/feedback cluster, the recurring mechanism was not replacement of teachers but iterative collaboration: AI generated options or feedback, learners evaluated and revised them, and teachers retained responsibility for task goals, calibration, and higher-order judgement. This pattern is consistent across controlled, qualitative, and mixed-methods studies, although longer-term transfer beyond the immediate intervention remains under-tested.

These benefits came with adoption and integrity caveats. Studies modelled learners' intention to use GenAI for L2 writing (Hu & Gong, 2025) and looked at trust in ChatGPT alongside perceived writing improvement (Anwer, 2026). Emotional engagement (Yao & Liu, 2026) and learner motivation toward ChatGPT versus instructor feedback (Barzanji & Alsofyani, 2026) also surfaced as factors in feedback uptake. At the instructional level, designs scaffolded academic writing (Godoy, 2026), brought GenAI into EFL classrooms through mixed methods (Chen et al., 2026), and worked to cultivate AI literacy and mediated learner autonomy (Hamamra et al., 2026). Case studies of transformed writing instruction (Duan & Wang, 2026), statement-of-purpose revision (Jiang et al., 2026), and EFL-major writing experiences (Meniado et al., 2024; Tsai et al., 2024) illustrated both the gains and the risks of over-reliance. Broader syntheses and enhancement studies echoed the pattern (Jaramillo et al., 2025; Yang et al., 2025). For RQ3, the answer is that GenAI most productively augments the writing process, supporting feedback, revision, and self-regulation, when learner agency and academic integrity are deliberately protected.

Comparative and Critical Analysis

Methodologically, short-duration quasi-experimental and qualitative designs dominate the corpus. Perception surveys and mixed-methods studies are common; fully randomised controlled trials are rare. Robust psychometric work, such as reliability studies of LLM scoring (Yavuz et al., 2025), sits alongside small-sample qualitative investigations (Huang & Wang, 2025), which makes for an uneven evidentiary base. Comparative feedback studies now lean more toward factorial and repeated-measures designs (Selvi et al., 2026; Tran, 2025), a sign of growing methodological sophistication. Even so, sample sizes tend to stay modest and follow-up periods short. Foreign-language writing-feedback research beyond English adds useful context to these trends (Carlini Versini et al., 2025).

Substantively, the field has moved on. Earlier work tended to show that GenAI could produce plausible language; current research asks how, for whom, and under what conditions it actually improves learning and assessment. That shift shows up in tool comparisons (Alpar, 2025), ESP applications (Jordán Enamorado, 2025), and writing implementations in Korea and Thailand (Lee & Jeong, 2026; Limna & Shaengchart, 2025; Tanharat & Phootirat, 2025). Still, discipline-specific evidence for applied sciences like sport sciences remains thin, and inconsistent outcome measures continue to limit cross-study comparability.

Taken as a whole, these findings put GenAI and LLMs in a consequential position within English-language education. Their strongest and most reliable effects appear when they are integrated into structured pedagogy and human-supervised assessment. Across learning, assessment, and writing, the evidence converges on a single point: the value of these tools rests less in autonomous performance and more in extending human instruction, amplifying feedback, personalising practice, and scaffolding revision.

Theoretically, these results stretch sociocultural and self-regulated learning accounts of language development. GenAI functions as a responsive mediating artefact, able to occupy the learner's zone of proximal development and offer just-in-time support that encourages agency and metacognition rather than passive

consumption. The collaborator framing that keeps recurring challenges author-centric models of writing, and it invites theorising about distributed, human, AI composition where authorship and cognition are shared yet still learner-directed.

On the practical side, there is actionable guidance here. In ESP contexts like sport sciences English, instructors can use GenAI to build discipline-specific vocabulary, rehearse professional speaking, and scaffold the drafting of abstracts and manuscripts. Rubric-based LLM scoring, meanwhile, belongs in formative rather than summative assessment. Curriculum designers should couple GenAI use with explicit feedback literacy and AI literacy instruction so that agency and integrity stay protected.

Comparison with prior reviews

The present synthesis is consistent with, but more narrowly organised than, a growing review literature. Balcı (2024) and Lo et al. (2024) both reported broad benefits of ChatGPT in EFL, while also identifying concerns about accuracy, integrity, and limited evidence outside writing. Qian (2025) placed GenAI within higher-education pedagogy and highlighted creativity, critical thinking, learner autonomy, and prompt literacy, with overreliance as a recurring risk. Albedah (2025) broadened the frame to multilingual AI and LLM applications, emphasising feedback, accessibility, bias, and the shortage of longitudinal evidence. Ebrahimi et al. (2025) similarly found that teacher-facing uses of ChatGPT were most valuable when mediated by teacher judgement and AI literacy. Karagoz (2025) and Parra Núñez and Castrillo de Larreta-Azelain (2025) converged on the value of AI-generated writing feedback for surface-level accuracy and revision, while retaining a strong role for human oversight. Mali (2025) emphasised ethical classroom practices in EFL/ESL writing, and Teng (2025) likewise framed ChatGPT as a useful but non-autonomous writing aid. Slamet and Basthomi (2025) highlighted contextual sensitivity, language accuracy, and the balance between AI and human interaction. Cong-Lem et al. (2025) added an equity lens, showing that access, AI literacy, bias, and infrastructure can shape who benefits. Li et al. (2026) moved the discussion toward task-based pedagogy and reported positive pooled effects with moderators including AI literacy, task planning, and assessment arrangements. Jeon (2025), in a review of reviews, synthesised 14 recent reviews/meta-analyses and argued for pedagogy-first, ethically grounded implementation. The current review extends this body of work by placing learning, assessment, and academic writing in one analytical map and by explicitly translating the evidence into ESP implications for sport sciences rather than assuming that general EFL findings automatically transfer.

Comparative contradictions and study-level explanations

Apparent contradictions are largely explainable by differences in design and outcome definition. For example, Abdelhalim and Alsehibany (2025) used a quasi-experimental classroom design with 71 learners and reported vocabulary gains, whereas Huang and Wang (2025) used a qualitative nine-learner writing study and emphasised selective uptake of AI suggestions. Selvi et al. (2026) used an 84-student, 10-week three-group quasi-experiment and reported benefits for writing and self-regulated learning, while Tran (2025) used a small 14-student experimental sequence study and showed that teacher-plus-AI sequencing mattered. Assessment evidence presents a similar distinction: Yavuz et al. (2025) reported very high reliability (ICC = 0.97), whereas Pack et al. (2024) showed that model-to-model reliability and validity were not identical. These studies do not provide a single effect estimate because their constructs, samples, durations, and outcome metrics differ. The appropriate synthesis is therefore conditional: GenAI appears most effective when tasks are structured, feedback is mediated, and outcomes are measured beyond immediate perception.

Limitations of this review

Several limitations constrain interpretation. First, the search used Scopus alone, so studies indexed only in Web of Science, ERIC, ProQuest, Education Source, or regional databases may have been missed. Second, only English-language journal articles were included, introducing language and publication bias. Third, screening was conducted by one reviewer and no independent duplicate-screening dataset was retained; consequently, Cohen's kappa and percentage agreement cannot be reported without inventing data. Fourth, the FICO rubric is a transparent reporting/applicability framework rather than a validated risk-of-bias instrument, and it cannot replace design-specific appraisal. Fifth, outcome measures were heterogeneous, ranging from vocabulary tests and speaking performance to essay scores, self-regulation scales, and qualitative reflections. Sixth, many interventions were short and some samples were small, limiting claims about long-term transfer. Finally, GenAI systems evolve rapidly, so model-specific findings may have limited temporal stability. Publication bias is also possible because the review did not search grey literature and most included studies reported some benefit or feasibility signal. These limitations mean that the findings should guide research and formative practice rather than be interpreted as definitive estimates of causal effect.

Operational future-study standards

Future studies should replace the phrase 'adequately powered' with transparent design targets. For two-group continuous outcomes, a practical minimum benchmark is 80% power at alpha = .05 for a prespecified primary

outcome; where a moderate standardized difference is the scientific target, Cohen's d = 0.50 can be used for planning rather than as a post hoc threshold. However, sample size should be calculated from the study's own primary outcome, expected attrition, and design. For psychometric or automated-assessment studies, researchers should report confidence intervals for reliability coefficients and criterion-related validity coefficients, not reliability alone. 'Discipline-specific' should mean that the corpus, tasks, rubrics, and vocabulary are anchored to identifiable sport-science genres such as research abstracts, lab reports, athlete case reports, exercise-prescription explanations, or professional coaching communication. Future intervention studies should preregister the primary outcome, use a comparator, report retention or delayed outcomes when feasible, and distinguish statistical significance from educationally meaningful change. These are recommendations for future study design, not thresholds retrospectively imposed on the heterogeneous studies synthesized here.

The evidence gaps are therefore substantive: discipline-specific ESP research in sport and health sciences is still sparse; longitudinal retention and transfer are limited; equity, access, integrity, and teacher professional development remain under-theorised; and common outcome metrics are not yet established across studies. These gaps reinforce the need for replication across institutions and contexts rather than further accumulation of short demonstrations.

A useful research agenda follows directly from the synthesis: preregistered longitudinal and controlled studies; discipline-specific corpora and validated rubrics for sport-science English; transparent reporting of prompts, model versions, intervention duration, and teacher mediation; independent replication; and explicit monitoring of learner agency, integrity, equity, and AI literacy.

In short, the evidence answers RQ1 by showing that GenAI supports skill development, most reliably when pedagogical scaffolding is in place. For RQ2, fine-tuned and GPT-4-class LLMs achieve high scoring reliability, though valid use demands human oversight. For RQ3, GenAI augments academic writing most productively through feedback, revision, and self-regulation, provided learner agency and integrity are protected.

Conclusions

This systematic review synthesised 50 peer-reviewed studies published from 2024 to 2026 on LLM and generative-AI applications in English-language learning, assessment, and academic writing. For RQ1, the evidence indicates that GenAI can support English-language development, especially vocabulary, speaking, and writing, but the clearest benefits occur in structured, teacher-mediated tasks and are not uniformly demonstrated across all skills or through long-term transfer measures. For RQ2, contemporary LLMs can show high scoring reliability in some rubric-based assessment tasks, but reliability should not be equated with validity, fairness, or readiness for high-stakes autonomous assessment. For RQ3, the most consistent role of GenAI is collaborative: feedback, revision, metacognitive support, and self-regulated learning are strengthened when learners critically evaluate AI output and teachers remain involved. The main contribution is an integrated evidence map together with an explicit translation rule for sport-science ESP: general EFL findings should be treated as transferable hypotheses until they are tested with sport-specific corpora, genres, and rubrics. The review is limited by a single Scopus database, English-only inclusion, single-reviewer screening, heterogeneous outcomes, and the rapid evolution of GenAI systems. Future research should prioritise 80%-powered preregistered studies, discipline-specific evaluation, external replication, and responsible AI implementation.

Acknowledgments

The Authors would like to thank Faculty of Sport Sciences, Universitas Negeri Padang for providing excellent facilities, as well as for being very supportive of this research from the beginning. Authors also thank the team members who had generously shared their brilliant ideas, engaging discussion, and academic supports during the study.

References

Abdelhalim, S. M., & Alsehibany, R. A. (2025). Integrating ChatGPT for vocabulary learning and retention: A classroom-based study of Saudi EFL learners. Language Learning and Technology, 29(1). https://doi.org/10.64152/10125/73635

Alasal, M. S. (2026). The impact of AI chatbots (ChatGPT) on EFL learners' writing fluency and accuracy. E-Learning and Digital Media. https://doi.org/10.1177/20427530261449472

Albedah, F. (2025). Artificial intelligence in language education: A systematic review of multilingual applications, large language models, and emerging challenges. Language Teaching Research Quarterly, 49, 247–268. https://doi.org/10.32038/ltrq.2025.49.13

Almarwani, M., & Alhenaky, R. (2026). Exploring Saudi EFL Learners' Perceptions of ChatGPT for Vocabulary Learning: A TAM-Based Descriptive Study. Arab World English Journal, 17(12), 147–162. https://doi.org/10.24093/awej/call12.9

Almutairi, H., Alfaifi, A. A., & Saleem, M. (2026). Writing Accuracy: How AI-Assisted Writing Instruction Can Support EFL Undergraduate Students. Information (Switzerland), 17(2), Article 157. https://doi.org/10.3390/info17020157

Alpar, Ö. (2025). Evaluating Generative AI Tools for Improving English Writing Skills: A Preliminary Comparison of ChatGPT-4, Google Gemini, and Microsoft Copilot. European Journal of Educational Research, 14(4), 1291–1308. https://doi.org/10.12973/eu-jer.14.4.1291

Alsofyani, A. H., & Barzanji, A. M. (2025). 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

Alyami, N. (2026). Investigating ChatGPT-mediated mind mapping to facilitate EFL learners' reading comprehension. PLOS ONE, 21(5 May), Article e0336185. https://doi.org/10.1371/journal.pone.0336185

Anwer, B. (2026). Trust in ChatGPT and Perceived Academic Writing Improvement: A TAM-based Quantitative Study in a ESL Context. International Journal of Technology in Education and Science, 10(1), 162–177. https://doi.org/10.46328/ijtes.5282

Balcı, Ö. (2024). The role of ChatGPT in English as a foreign language (EFL) learning and teaching: A systematic review. International Journal of Current Educational Studies, 3(1), 66–82. https://doi.org/10.46328/ijces.107

Barrios-Beltran, D. (2025). Exploring the efficacy of ChatGPT-4 feedback in second language Spanish writing. System, 133, Article 103771. https://doi.org/10.1016/j.system.2025.103771

Barzanji, A. M., & Alsofyani, A. H. (2026). Exploring Saudi EFL Learners' Motivation and Perceptions of ChatGPT and Instructor Feedback in Paragraph Writing. Arab World English Journal, 2026-January (Special Issue 3), 5–25. https://doi.org/10.24093/awej/AI3.1

Behforouz, B., Al Maqbali, A., Al Ghaithi, A., & Poorghorban, A. (2025). Mobile Interaction Meets AI Tutoring: Using ChatGPT-4o to Boost Speaking Skills in EFL Classrooms. International Journal of Interactive Mobile Technologies, 19(22), 34–49. https://doi.org/10.3991/ijim.v19i22.57447

Carlini Versini, D., Huang, J., Polo-Perez, N., & Zaher, A. (2025). Developing writing skills and feedback in foreign language education with chatGPT: a multilingual perspective. Language Learning in Higher Education, 15(2), 463–483. https://doi.org/10.1515/cercles-2024-0102

Chen, H., Rasool, U., Wang, R., & Dong, J. (2026). Integrating generative AI in EFL classrooms: A mixed-methods study on argumentative writing performance, writing enjoyment and motivation. Acta Psychologica, 268, Article 107216. https://doi.org/10.1016/j.actpsy.2026.107216

Chon, Y. V., & Shin, D. (2026). AI Chatbot as a Revision Aid in Second Language Writing: From Error Correction to Lexical Sophistication. Journal of Computer Assisted Learning, 42(3), Article e70259. https://doi.org/10.1002/jcal.70259

Cong-Lem, N., Bui, C. N. V., Nguyen, N. P. N., & Huynh, M. Q. (2025). Bridging or breaking? A systematic review of how generative AI shapes equity in foreign language education. International Journal of Applied Linguistics. https://doi.org/10.1111/ijal.70052

Duan, C., & Wang, Y. (2026). Transforming EFL writing instruction with generative AI: Evaluating feedback quality and learner independence. International Journal of Mechanical Engineering Education. https://doi.org/10.1177/03064190261445582

Ebrahimi, Z., Shakib Kotamjani, S., Qosimov, A., & Xodabande, I. (2025). Mapping the emerging research landscape on applications of ChatGPT in language teacher education: A systematic narrative literature review. Discover Artificial Intelligence, 5, 349. https://doi.org/10.1007/s44163-025-00631-z

Ennadir, H., & Habiballah, S. (2025). A Comparative Study of Human and AI-Assisted Assessment using ChatGPT: The Case of Moroccan EFL Learners. Arab World English Journal, 16(3), 175–187. https://doi.org/10.24093/awej/vol16no3.10

Franjić, M., & Božinovski, B. (2025). EXPLORING CHATGPT'S IMPACT on ESL WRITING PROFICIENCY: An EXPERIMENTAL APPROACH. Teaching English with Technology, 2025(2), 53–81. https://doi.org/10.56297/vaca6841/ZDDD4208/LUZG7476

Godoy, D. (2026). AI mediated scaffolding for second language academic writing in EFL composition classrooms. Discover Education, 5(1), Article 475. https://doi.org/10.1007/s44217-026-01386-0

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

Hamamra, B., N. Khlaif, Z., & Mahamid, N. (2026). AI literacy and mediated learner autonomy in ChatGPT-supported EFL writing: a qualitative study at a Palestinian university. International Journal of Educational Technology in Higher Education, 23(1), Article 19. https://doi.org/10.1186/s41239-026-00597-7

Hu, X., & Gong, W. (2025). Modeling Chinese EFL learners' intention to use generative AI for L2 writing through an integrated model of the TAM and TTF. Education and Information Technologies, 30(13), 18157–18179. https://doi.org/10.1007/s10639-025-13505-9

Huang, Y., & Wang, D. (2025). Can ChatGPT serve as a writing collaborator? Insights from Chinese EFL learners. System, 133, Article 103775. https://doi.org/10.1016/j.system.2025.103775

Jaashan, H. M. S., & Alashabi, A. A. (2025). Using AI Large Language Model (LLM-ChatGPT) to Mitigate Spelling Errors of EFL Learners. Forum for Linguistic Studies, 7(3), 328–339. https://doi.org/10.30564/fls.v7i3.8438

Jaramillo, J. J., Chiappe, A., & Delgado, F. S. (2025). From Struggle to Mastery: AI-Powered Writing Skills in ESL Education. Applied Sciences (Switzerland), 15(14), Article 8079. https://doi.org/10.3390/app15148079

Jeon, E. Y. (2025). Artificial Intelligence in ESL/EFL Education: Evidence from Recent Reviews (2024–2025). International Journal of Learning, Teaching and Educational Research, 24(10), 509–526. https://doi.org/10.26803/ijlter.24.10.24

Jiang, Y., Wu, Q., Yang, Y., Jian, C., & Zhao, J. (2026). Learner agency in revising GenAI-generated statements of purpose. British Journal of Educational Technology, 57(4), 965–983. https://doi.org/10.1111/bjet.70041

Jordán Enamorado, M. Á. (2025). Development of Linguistic Competence in English for Specific Purposes Through ChatGPT: A Case Study. Journal of Language and Education, 11(1), 85–100. https://doi.org/10.17323/jle.2025.23745

Karagoz, I. (2025). AI-generated feedback in English writing instruction for language learners: A systematic review. The Reading Matrix: An International Online Journal, 25(1), 51–67.

Kim, H.-S. (2025). AI-supported language learning for developing English listening and speaking skills and ChatGPT literacy. Korean Journal of English Language and Linguistics, 25, 1353–1377. https://doi.org/10.15738/kjell.25.202510.1353

Kohnke, L., Zou, D., & Su, F. (2025). Exploring the potential of GenAI for personalised English teaching: Learners' experiences and perceptions. Computers and Education: Artificial Intelligence, 8, Article 100371. https://doi.org/10.1016/j.caeai.2025.100371

Lai, C., Pan, M., Guo, K., & Cui, Y. (2026). What task to offload to GenAI in the writing feedback process? – Effects of task-offloading approaches on EFL learners' writing skill development. Computers and Education, 252, Article 105675. https://doi.org/10.1016/j.compedu.2026.105675

Lan, G., Li, Y., Yang, J., & He, X. (2025). Investigating a customized generative AI chatbot for automated essay scoring in a disciplinary writing task. Assessing Writing, 66, Article 100959. https://doi.org/10.1016/j.asw.2025.100959

Lee, Y. J., & Jeong, S. H. (2026). Implementing AI-integrated English writing: university English language learners writing strategies and perceptions. International Journal of Evaluation and Research in Education, 15(1), 805–814. https://doi.org/10.11591/ijere.v15i1.35461

Li, Y., Shadiev, R., & Chiu, T. K. F. (2026). Applying generative artificial intelligence to task-based language teaching and learning: A systematic review and meta-analysis. TechTrends, 70, 150–163. https://doi.org/10.1007/s11528-025-01140-7

Limna, P., & Shaengchart, Y. (2025). Generative AI as an English Writing Aid: Thai University Students' Perceptions and Experiences with ChatGPT and Gemini. Suranaree Journal of Social Science, 19(3), Article e279466. https://doi.org/10.55766/sjss279466

Lo, C. K., Yu, P. L. H., Xu, S., Ng, D. T. K., & Jong, M. S. Y. (2024). Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies. Smart Learning Environments, 11, 50. https://doi.org/10.1186/s40561-024-00342-5

Lu, C. T., Chen, Y. J., Wu, T. Y., & Lu, Y. Y. (2026). An Intelligent English-Speaking Training System Using Generative AI and Speech Recognition. Applied Sciences (Switzerland), 16(1), Article 189. https://doi.org/10.3390/app16010189

Maleki, A. (2026). Effects of AI generated and teacher feedback on EFL learners writing performance and emotional experience. Discover Artificial Intelligence, 6(1), Article 199. https://doi.org/10.1007/s44163-026-00935-8

Mali, Y. C. G. (2025). Exploring the use of ChatGPT in EFL/ESL writing classrooms: A systematic literature review. Journal of Language and Education, 11(2), 137–156. https://doi.org/10.17323/jle.2025.21793

Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: a study on proficiency, revision frequency and writing quality. Discover Education, 4(1), Article 170. https://doi.org/10.1007/s44217-025-00602-7

Meniado, J. C., Huyen, D. T. T., Panyadilokpong, N., & Lertkomolwit, P. (2024). Using ChatGPT for second language writing: Experiences and perceptions of EFL learners in Thailand and Vietnam. Computers and Education: Artificial Intelligence, 7, Article 100313. https://doi.org/10.1016/j.caeai.2024.100313

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence, 6, Article 100234. https://doi.org/10.1016/j.caeai.2024.100234

Parra Núñez, R., & Castrillo de Larreta-Azelain, M. D. (2025). The impact of using ChatGPT for feedback in EFL writing: A systematic review. Revista Internacional de Lenguas Extranjeras, 23, 95–119. https://doi.org/10.17345/rile23.4221

Pretorius, M., & Thewissen, J. (2026). Mediating L2 writing revision: The impact of teacher-moderated ChatGPT feedback on L2 accuracy and lexical diversity. Journal of Second Language Writing, 71, Article 101286. https://doi.org/10.1016/j.jslw.2026.101286

Qian, Y. (2025). Pedagogical applications of generative AI in higher education: A systematic review of the field. TechTrends, 69, 1105–1120. https://doi.org/10.1007/s11528-025-01100-1

Selvi, A. E., Irgatoğlu, A., Yurttadur, O., & Dağbaşı, G. (2026). Sustainable L2 writing pedagogy in Turkish higher education: Effects of AI-mediated feedback on self-regulated learning and writing performance. PloS one, 21(7), e0344618. https://doi.org/10.1371/journal.pone.0344618

Shahzad, W., & Abbas, F. (2026). Automated Evaluation of ESL Learners' English Writing Skills in English-Medium Instruction (EMI) through AI Writing Analytics. Journal of Communication, Language and Culture, 6(1), 147–166. https://doi.org/10.33093/jclc.2026.6.1.8

Slamet, J., & Basthomi, Y. (2025). Examining the challenges and opportunities of ChatGPT in EFL education: A systematic literature review. Journal of University Teaching and Learning Practice, 22(2). https://doi.org/10.53761/deezkh88

Tai, T. Y., & Chen, H. H. J. (2024). Navigating elementary EFL speaking skills with generative AI chatbots: Exploring individual and paired interactions. Computers and Education, 220, Article 105112. https://doi.org/10.1016/j.compedu.2024.105112

Tanharat, N., & Phootirat, P. (2025). Integrating Generative AI in EFL Academic Writing: Thai English-Major Students' Purposes, Perceptions, and Experiences with ChatGPT. rEFLections, 32(3), 1891–1916. https://doi.org/10.61508/refl.v32i3.285982

Teng, M. F. (2025). 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

Tram, N. H. M., Nguyen, T. T., & Tran, C. D. (2024). ChatGPT as a tool for self-learning English among EFL learners: A multi-methods study. System, 127, Article 103528. https://doi.org/10.1016/j.system.2024.103528

Tran, T. T. T. (2025). Enhancing EFL Writing Revision Practices: The Impact of AI- and Teacher-Generated Feedback and Their Sequences. Education Sciences, 15(2), Article 232. https://doi.org/10.3390/educsci15020232

Tsai, C. Y., Lin, Y. T., & Brown, I. K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. Education and Information Technologies, 29(17), 22427–22445. https://doi.org/10.1007/s10639-024-12722-y

Yang, N., Gao, N., & Zhang, T. (2025). Artificial Intelligence in EFL Writing Enhancement: A Study on Chatbot Feedback and Learner Autonomy. International Journal of Web-Based Learning and Teaching Technologies, 20(1). https://doi.org/10.4018/IJWLTT.390442

Yao, G., & Liu, Z. (2026). ChatGPT feedback and emotional engagement in L2 writing: A control-value theory perspective using Q-methodology. Assessing Writing, 68, Article 101045. https://doi.org/10.1016/j.asw.2026.101045

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(1), 150–166. https://doi.org/10.1111/bjet.13494

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