REVIEW ARTICLE

Asian Research Association

Indian Journal of
LANGUAGE AND LINGUISTICS

DOI: 10.54392/ijll2627

Generative AI in L2 Writing, Feedback, and Assessment: A Systematic Critical Review of Validity, Voice, Equity, and Pedagogical Design

J. Charly Jerome [a, *]

[a] Department of Language and Communication, KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India
[*] Corresponding author Email: charlyjerome@kpriet.ac.in
DOI: https://doi.org/10.54392/ijll2627
Received: 07-03-2026; Revised: 11-06-2026; Accepted: 17-06-2026; Published: 28-06-2026

Abstract: Writing in a second language has never been a simple skill to teach or learn, and the arrival of generative artificial intelligence has made the challenge considerably more complicated. Large language models entered classrooms, feedback systems, and assessment platforms so quickly that research could not keep pace with practice. Teachers began revising integrity policies before anyone had studied whether AI use was helping students grow as writers. Assessment bodies began questioning whether their rubrics still measured what they were designed to measure. This systematic critical review brings together 74 empirical studies, systematic reviews, and conceptual papers published between January 2022 and April 2026, a period shaped decisively by the release of ChatGPT and related models. Searches were conducted across five major bibliographic databases using the PRISMA 2020 protocol. The review synthesises evidence on how generative AI tools support or disrupt the writing process, the feedback cycle, and the scoring of written texts in second and additional language contexts. Four tensions run through the entire body of scholarship: questions about what AI scoring systems actually measure, concerns about the erosion of writer voice and identity, unequal access across language backgrounds and resource settings, and the absence of principled frameworks that could help teachers use these tools thoughtfully. The review also documents recurring methodological problems in the current literature, including small samples, short observation windows, and thin theoretical grounding. A research agenda is proposed around construct validity in AI scoring, voice and authorship in human and AI collaborative writing, equity and multilingual access, and ethics in automated assessment at scale. The review offers teachers, institutions, and journal editors a conceptual map of the field, an honest critique of its current weaknesses, and concrete directions for the next phase of scholarship.

Keywords: Generative AI, L2 Writing, Automated Feedback, Writing Assessment, Equity, Pedagogical Design

1. Introduction

1.1 Why Generative AI in L2 Writing Needs a New Review

Second language writing has always attracted technology. Spell checkers, grammar correction tools, corpus assistants, and automated scoring systems each arrived in classrooms with promises and complications. Each wave forced teachers and researchers to ask the same basic questions: what does writing actually require of a person, who gets to call themselves a writer, and what kind of feedback genuinely helps a learner grow. Generative artificial intelligence does not simply continue this pattern. It represents a different kind of change altogether (Krakowski, 2025). Large language models can now produce several paragraphs of fluent, grammatically polished writing from a short prompt. They can respond to a student draft with detailed suggestions that look very much like a trained teacher's feedback. They can score essays with a consistency that makes human raters look unreliable by comparison. None of this happened gradually. It arrived almost overnight, and the second language writing research community is still working out what it means.

The speed of adoption has been striking. Within a few months of ChatGPT becoming publicly available in late 2022, students at universities across Asia, Africa, Europe, and the Americas were using AI tools to brainstorm,

draft, and revise their writing (Li, 2026). Teachers found themselves in classrooms where the boundary between a student's own text and an AI generated one was genuinely unclear. Institutions scrambled to revise their academic integrity policies without any solid research evidence about whether AI use was affecting learning outcomes one way or the other. Assessment developers faced urgent questions about whether their rubrics and scoring procedures remained valid in a world where AI could consistently produce text that scored at the top of the scale. These are not hypothetical problems. They are the daily reality of L2 writing instruction as it stands today (Zhang, *et al.,* 2025).

Earlier reviews of AI in language education exist, but they either predate the generative AI period or address only narrow aspects of it. Reviews of automated writing evaluation have typically treated the technology as a scoring mechanism rather than as something that participates actively in the writing process (Conijn *et al.,* 2020). Reviews of AI feedback have concentrated on corrective feedback in narrow grammatical domains without engaging with broader pedagogical or equity questions. No existing systematic review addresses writing support, feedback, and assessment together in a way that reveals how problems in one domain drive problems in the others. This field needs a review that honestly evaluates what is known, identifies where the evidence is weak, and points toward a more principled direction for research.

This review is organised around four conceptual tensions that run through the entire literature and cannot be resolved independently of one another. The first is between measurement efficiency and construct validity: AI scoring systems may be consistent, but consistency is not the same as measuring the right thing (Aryadoust *et al.,* 2020). The second is between writing support and the suppression of writer voice: AI tools may help learners produce more fluent text while displacing the linguistic risk-taking and identity expression that drives genuine language growth (Hakim, 2023). The third is between technological access and educational equity: AI writing tools are designed primarily for English speakers, and their benefits are distributed very unevenly across language backgrounds and resource contexts. The fourth is between what AI can do and what teachers should do with it in the classroom: the research has spent far more energy on the first question than on the second.

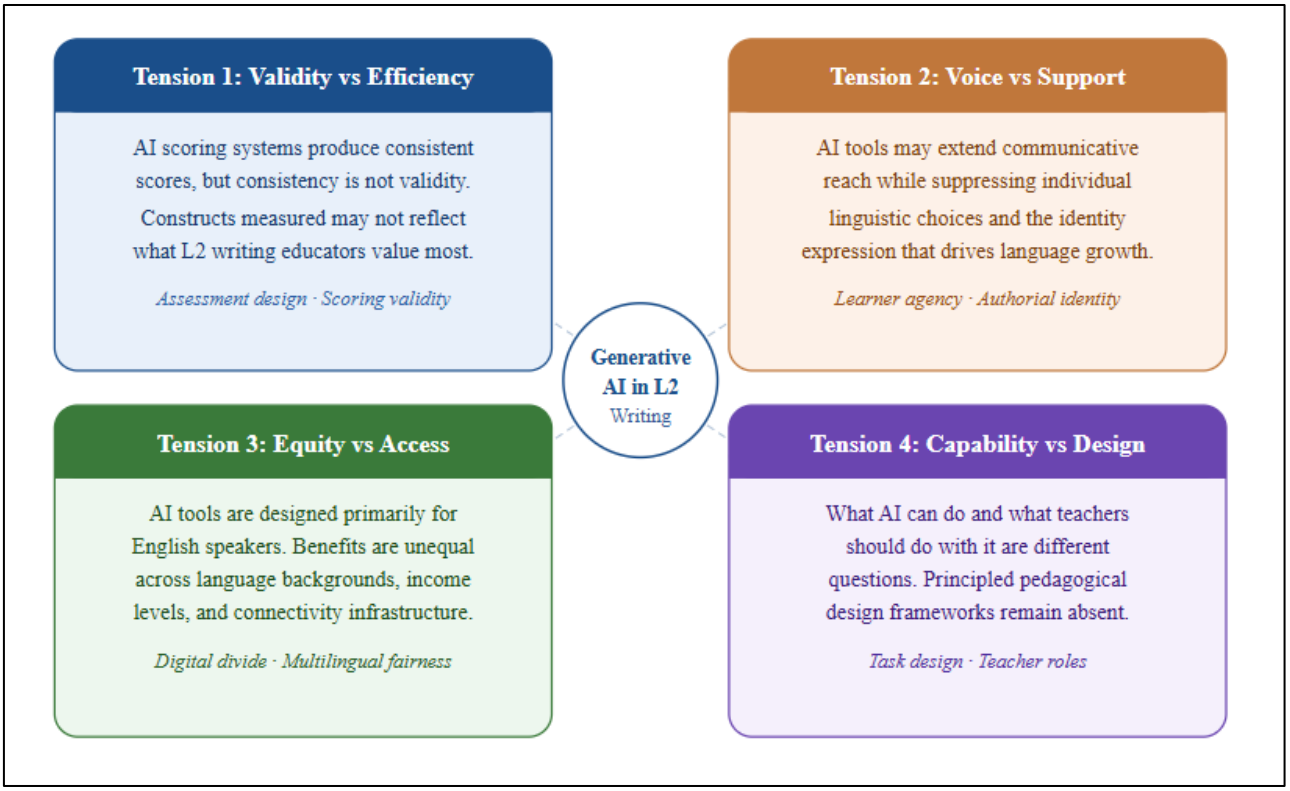

Figure 1. Conceptual framework: four cross-cutting tensions in generative AI and L2 writing research

This review has five main goals. First, it shows where empirical and conceptual research on generative AI stands right now in the areas of L2 writing, feedback, and assessment. Second, it critically assesses the assertions of validity and reliability regarding AI assessment tools in L2 contexts. Third, it looks at how using generative AI changes the writer's voice, the learner's agency, and the author's identity. Fourth, it looks into how different groups

of learners and schools can use and benefit from AI writing tools in different ways. Fifth, it puts forward a research agenda and practical suggestions based on the evidence that was looked at.

1.2 Conceptual Boundaries

1.2.1 Generative AI, Large Language Models and AI writing Support

Generative artificial intelligence refers to computational systems that produce new content text, images, code, audio that was not explicitly present in their training data. In the context of writing, the most relevant class of these systems is large language models. These are neural network systems trained on very large bodies of text, capable of producing fluent and contextually appropriate written output in response to prompts (Bender & Koller, 2020). ChatGPT, GPT-4, Gemini, Claude, and related systems belong to this category. What distinguishes these tools from earlier natural language processing systems is their generative capacity: they do not simply analyse or markup existing text, they actively produce new text. This is what makes them both educationally useful and pedagogically problematic. The conceptual framework image was given in figure 1.

AI writing support is a broader term that covers any AI based tool used to assist in the composition process. This ranges from grammar and style checkers like Grammarly, which uses machine learning to detect and suggest corrections in existing text, through writing assistants that offer suggestions for continuing or revising a draft, to fully generative systems that can produce complete essays from a single topic prompt (Yang *et al.,* 2025). The distinction matters for pedagogy because learner agency varies enormously across these tools. A grammar checker leaves the learner firmly in control of what they are writing and offers targeted, local suggestions. A fully generative model, if used without critical engagement, can effectively remove the learner from the writing process altogether. Research that treats all AI writing tools as equivalent obscures pedagogical differences that matter a great deal.

1.2.2 Automated Writing Evaluation, Automated Essay Scoring, Chatbot Assistance, and Prompt-Based Writing Support

Automated writing evaluation is an umbrella term for systems that use computational methods to provide feedback on student writing. It covers a wide range of tools, from early systems that counted words and parsed syntactic structures to contemporary systems built on transformer-based neural architectures (Mizumoto & Eguchi, 2023). Automated essay scoring is a sub-category focused specifically on assigning a numerical score or grade to a piece of writing. Systems such as e-rater, Turnitin's Writing Match, and various national assessment platforms have used automated scoring for decades, primarily in large-scale standardised testing contexts. These systems were developed before the generative AI period and operate on fundamentally different principles from large language models, though newer scoring approaches are beginning to incorporate large language model components.

Chatbot assistance in writing contexts refers to the use of conversational AI systems to support learners in the writing process through interactive dialogue. A learner might ask a chatbot to explain a grammar rule, suggest vocabulary alternatives, identify weaknesses in a draft, or help generate ideas for an essay (Wiboolyasarin, 2025). This differs from prompt-based writing support, where a learner submits a prompt and receives a generated text, because chatbot interaction is inherently dialogic and involves sustained back-and-forth exchange. The pedagogical value of chatbot assistance may be greater than that of simple prompt-and-response generation because it can approximate the kind of scaffolded dialogue that sociocultural theories of writing development emphasise. Whether it produces meaningfully different learning outcomes is a question the research has not yet settled.

1.3 Review Questions and Contribution

This systematic critical review is organised around four primary research questions. The first asks what the available evidence reveals about how generative AI tools support or disrupt the L2 writing process across different phases of composition, from brainstorming through to final editing. The second asks what the validity, reliability, and fairness properties of AI generated feedback and AI scoring systems are in L2 writing contexts, and what the evidence reveals about alignment with established constructs of writing quality. The third question asks: how does the use of generative AI in writing, feedback, and assessment affect L2 writers' sense of voice, agency, identity, and authorship? The fourth question asks: how are the benefits and risks of generative AI in L2 writing distributed across different

learner populations, language backgrounds, and institutional contexts, and what equity implications follow from current patterns of development and deployment?

This paper makes four contributions to the field. First, it provides a comprehensive synthesis of research published after the emergence of ChatGPT, covering a period that earlier reviews could not address. Second, it integrates three domains — writing support, feedback, and assessment — that the existing literature tends to treat separately, thereby revealing cross-cutting tensions that are invisible when each domain is reviewed in isolation. Third, it offers a sustained methodological critique of the current literature, identifying patterns of weakness that have important implications for how future studies should be designed. Fourth, it proposes a research agenda and a set of practical recommendations grounded in the evidence reviewed rather than in speculation about AI's potential. The contribution is both synthetic and critical: it takes stock of what is known, is honest about what remains uncertain, and offers a principled direction for the next phase of scholarship.

2. Methodology

2.1 Review Design and Epistemological Rationale

This study uses a systematic critical review design. This design combines the replicable and auditable procedures of systematic review methodology with the interpretive depth that a purely quantitative synthesis cannot provide. A meta-analysis would not be appropriate for this literature because the conditions for statistical pooling (shared outcome measures, comparable populations, and comparable intervention specifications) are not consistently met across the included studies (Page *et al.,* 2021a). The review questions extend beyond aggregating effect sizes to evaluating the theoretical and methodological assumptions that underpin the existing evidence base. This task requires interpretive judgement. The systematic critical review design preserves the rigour and transparency of systematic searching and screening while retaining the analytical capacity needed to produce a genuinely critical account of the field.

The review was conducted and reported in full compliance with the PRISMA 2020 statement (Page *et al.*, 2021b). PRISMA 2020 is the current international standard for reporting systematic reviews across disciplines. It updated the earlier PRISMA 2009 guidelines to address new developments in systematic review practice, including more detailed documentation of deduplication and screening procedures and explicit guidance on the treatment of preprints and grey literature (Page *et al.,* 2021a). The PRISMA flow diagram is presented in Figure 2. Although this review addresses social science and education research questions outside the registration scope of the PROSPERO registry, the search strategy, inclusion and exclusion criteria, and quality appraisal protocol were each finalised before the primary database searches were executed, preventing post-hoc modification of review parameters (Efthimiou *et al.,* 2016).

2.2 Search Sources and Database Rationale

Systematic literature searches were conducted across five major bibliographic databases: Web of Science Core Collection, Scopus, Education Resources Information Center (ERIC), Linguistics and Language Behaviour Abstracts (LLBA), and the MLA International Bibliography. These five sources were selected because their combined coverage encompasses the four disciplinary homes of the target literature: applied linguistics and second language acquisition, language testing and writing assessment, educational technology, and broader educational research. No single database provides adequate coverage across all four domains (Efthimiou, *et al.,* 2016). Web of Science Core Collection and Scopus were included as the two most comprehensive multidisciplinary databases available. ERIC was included to capture specialist education research not fully indexed by multidisciplinary databases. LLBA was included as the primary database for linguistics, providing deep indexing of journals in second language writing and language testing. The MLA International Bibliography was included for complementary coverage of language, literature, and rhetoric journals capturing humanistically oriented scholarship on writing and technology.

Google Scholar was used exclusively for citation chasing both backward chasing of references cited in included studies and forward chasing of papers that subsequently cited those studies; and was not used as a primary screening source (Page *et al.,* 2021a).

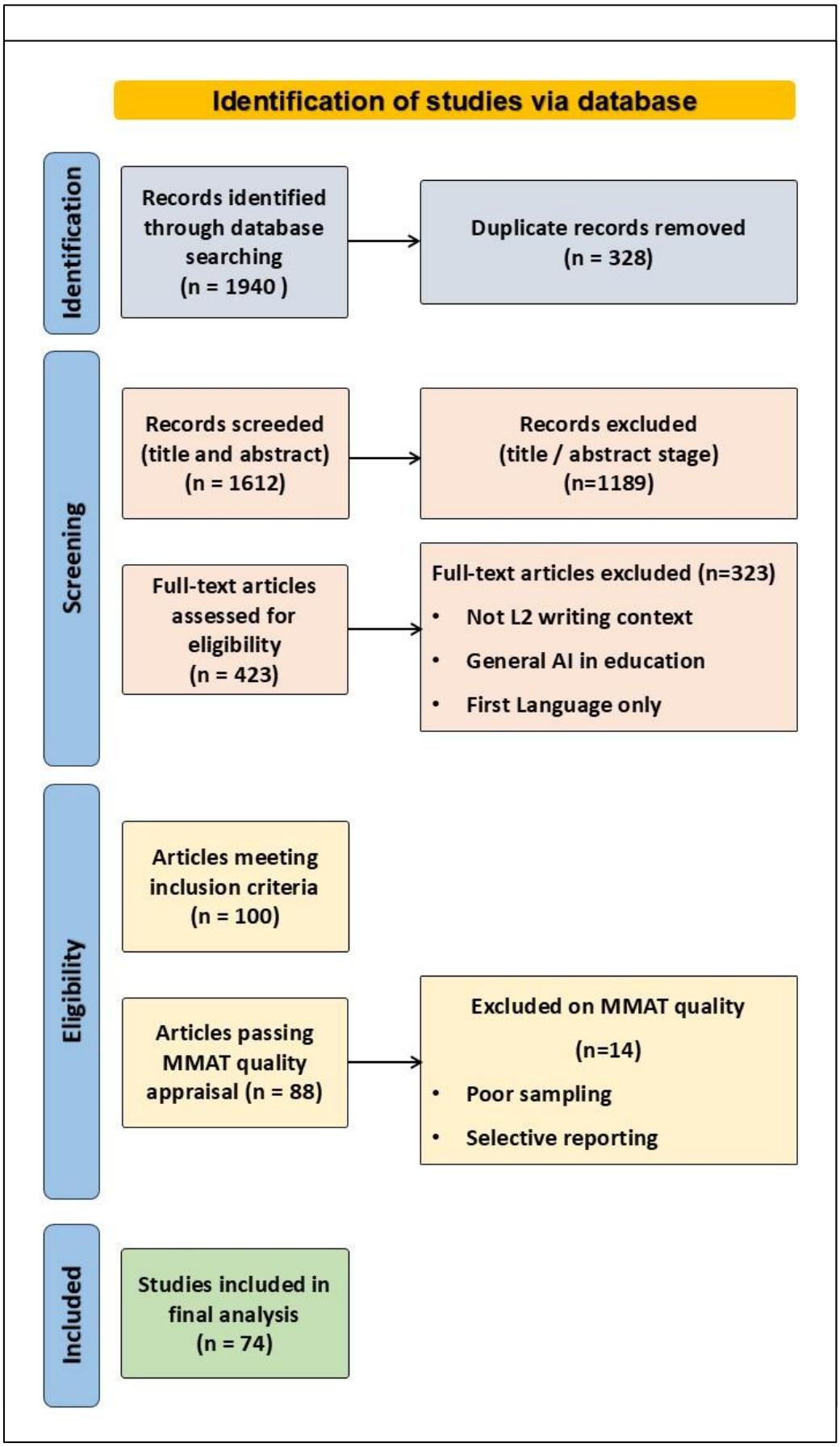

Figure 2. PRISMA 2020 flow diagram illustrating the identification, screening, eligibility assessment and inclusion of studies.

While Google Scholar offers unmatched breadth of coverage, its inconsistent indexing policies, inability to reliably deduplicate results at scale, and routine inclusion of documents that are not peer reviewed make it unsuitable as a primary source for a PRISMA-compliant review. Its use for citation chasing, however, is a legitimate and important supplement that can identify influential papers published in journals not indexed by the five primary databases (Efthimiou *et al.,* 2016).

2.3 Search Strategy and String Construction

The search strategy was built following the Population, Concept, Context (PCC) framework recommended by the Joanna Briggs Institute (Bour *et al.,* 2021). The population of interest was learners engaged in writing in a second, foreign, or additional language. The concept was the use of generative artificial intelligence tools for writing support, written feedback, or writing assessment. The context was any educational or assessment setting where L2 writing occurs. The following Boolean search string was applied across all five primary databases with adaptations to each database's field codes and syntax conventions: ("generative AI" OR "large language model*" OR LLM* OR ChatGPT OR GPT OR "artificial intelligence") AND ("L2 writing" OR "second language writing" OR "EFL writing" OR "ESL writing" OR "written feedback" OR "writing feedback" OR "writing assessment" OR "automated writing evaluation" OR "automated essay scoring"). In ERIC, supplementary controlled descriptor terms were added. The time window was set from January 2022 to the final search date in April 2026, reflecting the consensus that the post-ChatGPT literature constitutes a qualitatively distinct body of work (Sahan *et al.,* 2023).

2.4 Inclusion and Exclusion Criteria

Before any screening started, the criteria for inclusion and exclusion were clearly spelled out. This made it impossible to change the eligibility standards after the fact (Page *et al.,* 2021a). Studies were included only if they concurrently met all five of the following criteria. First, the study was published in a peer-reviewed journal that is listed in one of the five main databases or found through citation chasing. This does not include preprints, conference abstracts, or book chapters, no matter how good they seem. Second, the study was empirical, collecting and analysing primary or secondary data through any recognised research method, or a systematic review, or a major conceptual paper making a documented theoretical or methodological contribution to the field. Third, the study was conducted in a second, foreign, or additional language writing context, including EFL, ESL, English for Academic Purposes, and foreign language contexts for languages other than English. Fourth, the study examined the use of generative AI or related automated tools in relation to writing process support, written feedback, or writing assessment. Fifth, the study was published between January 2022 and the final search date.

Studies were excluded on six grounds (Table 1). Studies addressing AI in education in general without a specific and sustained focus on language writing were excluded, accounting for 98 of the 323 full-text exclusions. Studies focused exclusively on first language writing without a direct comparative element involving L2 writers were excluded. Studies in natural language processing or computational linguistics without pedagogical, assessment, or language development engagement were excluded. Studies published in journals identified as predatory using Beall's criteria and verified against the Directory of Open Access Journals were excluded (Richtig *et al.,* 2023). Items that were not peer reviewed, including editorials, opinion pieces, and preprints, were excluded. Duplicate publications reporting the same study were identified using Rayyan and cross-checked manually; only the most complete version was retained.

2.5 Screening Process and Inter-Rater Reliability

Screening was conducted in two sequential stages using Rayyan, a web-based systematic review management platform that supports blinded independent screening and structured consensus resolution (Ouzzani, 2016). In the first stage, two independent reviewers screened all 1,612 unique records on the basis of title and abstract alone, applying the five inclusion criteria and six exclusion criteria. Records classified as uncertain by either reviewer were automatically carried forward to full-text screening so that borderline cases were not lost. At the conclusion of title and abstract screening, 1,189 records were excluded and 423 were retained for full-text assessment. Inter-rater agreement was assessed using Cohen's kappa (Cohen, 1960), yielding $\kappa = 0.81$, which indicates strong agreement (Table 2).

Table 1. Inclusion and exclusion criteria with operational definitions and validation

Criterion Type	Criterion	Operational Definition	Validation
INCLUDE	Peer reviewed journal publication	Indexed in WoS, Scopus, ERIC, LLBA, or MLA, or retrieved via citation chasing	Ensures minimum quality assurance standard for evidence base
INCLUDE	Appropriate study design	Empirical (any method), systematic review, or major conceptual paper	Ensures verifiable evidence or theoretical contribution
INCLUDE	L2/FL writing population	Second, foreign, or additional language learners in any writing context	Defines the population of interest
INCLUDE	Relevance to focal domain	Focuses on writing process, AI feedback, or AI assessment in L2 writing	Ensures direct bearing on the review questions
INCLUDE	Time window	Published January 2022 to final search date (April 2026)	Captures post-generative AI literature
EXCLUDE	General AI in education	AI without specific sustained L2 writing focus	Outside focal domain; n=98 excluded at full-text stage
EXCLUDE	First-language writing only	No L2 comparative element	Outside population of interest; n=22 excluded
EXCLUDE	NLP/CS without pedagogical content	No engagement with learning, teaching, or assessment	Different disciplinary tradition; n=43 excluded
EXCLUDE	Non-peer-reviewed	Editorial, opinion, preprint, or conference abstract	Insufficient quality assurance; n=57 excluded
EXCLUDE	Predatory journal source	Identified via Beall's criteria or absent DOAJ quality flag	Data integrity cannot be verified; n=14 excluded

Table 2. Screening outcomes and inter-rater agreement statistics by stage.

Screening Stage	Records Entering	Records Excluded	Records Retained	Cohen's Kappa	Agreement Level
Title and abstract screening	1,612	1,189	423	$\kappa = 0.81$	Strong
Full-text eligibility screening	423	323	100	$\kappa = 0.84$	Strong
MMAT quality appraisal	100	26	74	ICC = 0.79	Acceptable

Table 3. Reasons for full-text exclusion with frequencies

Exclusion Reason	Number of Studies (n)	Percentage of Full-Text Exclusions (%)
Not L2/EFL/ESL writing context — general education AI study	98	30.3
AI in education without L2 writing focus as primary domain	89	27.6
Non-peer-reviewed: opinion, editorial, or preprint	57	17.6
NLP/computational linguistics — no pedagogical relevance	43	13.3
Exclusively first-language writing, no L2 comparative element	22	6.8
Predatory or non-credibly peer-reviewed journal source	14	4.3
Total	323	100.0

In the second stage, full-text copies of all 423 retained records were retrieved and independently screened by both reviewers (Table 3). A total of 323 studies were excluded at the full-text stage. The most common reason for exclusion was absence of a specific L2 writing focus within an otherwise AI-and-education study (n = 98), followed by general AI-in-education papers without sufficient engagement with writing as a distinct construct (n = 89), sources that were not peer reviewed (n = 57), natural language processing or computer science papers without pedagogical relevance (n = 43), exclusively first-language writing without a comparative L2 element (n = 22), and predatory or insufficiently credentialed journal sources (n = 14). Following full-text screening, 100 studies remained eligible for quality appraisal. Inter-rater agreement at the full-text stage was κ = 0.84.

2.6 Quality Appraisal Using the Mixed Methods Appraisal Tool

All 100 studies that passed two-stage eligibility screening were subjected to formal quality appraisal using the Mixed Methods Appraisal Tool Version 2018 (MMAT 2018), (Hong, 2018). MMAT 2018 was selected because it is specifically designed for literature containing qualitative, quantitative, and mixed methods studies, which is precisely the methodological composition of the target body of research. For each study type, five binary quality criteria are applied, yielding a quality score out of five. Quality appraisal was conducted independently by two reviewers, with a training session held before appraisal began using five pilot studies not in the included set. Inter-rater agreement on MMAT scores was assessed using the intraclass correlation coefficient, yielding ICC = 0.79, which meets the threshold of 0.75 recommended for acceptable agreement. Disagreements were resolved through structured discussion; in three cases where consensus was not reached, the senior reviewer adjudicated.

Studies scoring three or more out of five on the MMAT criteria were retained for the primary synthesis, consistent with published guidance for mixed methods systematic reviews in education research. Applying this threshold resulted in the exclusion of 26 studies from the 100 that reached quality appraisal: 14 were excluded because of inadequate description of sampling procedures and participant characteristics; seven because of absent or inadequate analysis of potential confounds; three because of selective or incoherent reporting of findings; and two because the qualitative strand lacked adequate analytical rigour. The 26 excluded studies were not simply discarded. Their systematic documentation of methodological weaknesses contributes directly to the methodological critique presented in Section 10 of this review. The final included sample therefore comprised 74 studies as listed in table 4.

Table 4. Final list of selected studies for analysis

No.	Authors & Year	Country	Design	AI Tool	Domain	Key Finding Summary
1	Yan (2023)	China	Qualitative	ChatGPT	Writing process	ChatGPT supported L2 writing at all stages. Students valued pedagogical utility but raised concerns about plagiarism, equity, and academic integrity.
2	Su *et al*. (2023)	Taiwan	Mixed methods	ChatGPT	Writing process / Feedback	Collaborative use of ChatGPT in argumentative writing improved content quality. Dialogic engagement with AI promoted critical thinking in revision.
3	Barrot (2023)	Philippines	Conceptual review	ChatGPT	Writing process / Assessment	Identified key pitfalls (over-reliance, voice loss, integrity risks) and potentials (formative feedback, idea generation) of ChatGPT for L2 writing.

4	Zou & Huang (2024)	China	Qualitative	ChatGPT	Writing process / Voice	Doctoral students perceived ChatGPT useful at pre-, during-, and post-writing stages but worried about learning loss and authorial voice erosion.
5	Song & Song (2023)	China	Mixed methods	ChatGPT	Writing process / Motivation	ChatGPT improved EFL students' academic writing skills and motivation. Qualitative data confirmed positive affective engagement throughout writing tasks.
6	Yang & Li (2024)	China	Systematic review	ChatGPT / LLMs	Writing process	Synthesised ChatGPT roles in L2 learning; found gains in writing, vocabulary, and grammar but persistent risks of over-reliance and ethical concerns.
7	Kohnke *et al.* (2023)	Hong Kong	Conceptual	ChatGPT	Writing process / Teaching	Argued ChatGPT can scaffold L2 writing across brainstorming, drafting, and revision if teachers provide structured prompts and reflection activities.
8	Godwin-Jones (2024)	USA	Conceptual / review	ChatGPT / LLMs	Writing process / Agency	Theorised distributed agency in L2 writing with AI; argued SLA will shift as AI provides formative interaction, fundamentally changing the teacher's role.
9	Mahapatra (2024)	India	Mixed methods	ChatGPT	Writing process / Feedback	ChatGPT as formative feedback tool significantly improved ESL undergraduates' academic writing skills. Student perceptions were overwhelmingly positive.
10	Xiao & Zhi (2023)	China	Survey / qualitative	ChatGPT	Writing process	EFL learners used ChatGPT primarily for grammar checking and vocabulary selection while consciously attempting to maintain their own writing voice.
11	Wang (2024)	USA	Qualitative	ChatGPT	Writing process / Voice	Native and non-native English speakers differed in AI-assisted writing processes; NNS relied more

						on AI for language accuracy, NS for idea elaboration.
12	Wang & Wang (2025)	USA	Qualitative case study	ChatGPT	Writing process / AI literacy	Proposed the APSE model of critical AI literacy for L2 writers. Students used ChatGPT across brainstorming, outlining, revising, and editing stages.
13	Escalante *et al.* (2023)	USA	Quasi-experimental	GPT-4	Feedback	AI-generated writing feedback produced learning outcomes equivalent to human feedback. Students were evenly divided between preferring AI or teacher feedback.
14	Guo & Wang (2024)	China	Comparative / Mixed	ChatGPT	Feedback	ChatGPT provided more voluminous and balanced feedback than teachers. Teacher feedback was selective and content-focused. A complementary combination was recommended.
15	Pack *et al.* (2024)	USA	Longitudinal / Quantitative	GPT-4 / Claude / PaLM	Assessment / Feedback	GPT-4 showed best validity and reliability for AES of ELL writing. All LLMs improved in intrarater reliability over time; GPT-3.5 was least consistent.
16	*Teng (2024)*	China	Quantitative (n=174)	ChatGPT	Feedback	Compared peer feedback and ChatGPT feedback in EFL writing; ChatGPT group showed superior writing performance and higher engagement with feedback.
17	Liu *et al.* (2024)	China	Quasi-experimental	LLM-based system	Writing / Self-regulation	Integrating LLMs into EFL instruction improved writing performance and self-regulated learning strategies. Effects were stronger for lower-proficiency learners.
18	Koltovskaia *et al.* (2024)	USA	Mixed methods case study	ChatGPT	Feedback engagement	Graduate students accepted over 60% of ChatGPT feedback on academic text revision. High cognitive engagement was observed when learners identified inaccurate AI feedback.

19	Lu *et al.* (2024)	China	Mixed methods	ChatGPT	Feedback / Assessment	ChatGPT effectively marked undergraduate writing but provided more general suggestions. Teacher comments were more specific and emotionally engaging for students.
20	Steiss *et al.* (2024)	USA	Quasi-experimental	ChatGPT	Feedback	EFL learners receiving iterative ChatGPT feedback over multiple rounds showed greater writing improvement than those receiving a single round of AI feedback.
21	Lin & Crosthwaite (2024)	Hong Kong	Comparative analysis	ChatGPT	Feedback	Teacher feedback emphasised indirect prompts on content; ChatGPT feedback was more direct and distributed more evenly across language, organisation, and content.
22	Long (2024)	Hong Kong	Action research	ChatGPT	Corrective feedback	ChatGPT as written corrective feedback tool improved grammatical accuracy in EFL. Students initially needed scaffolding to interact dialogically with the AI tool.
23	Lan (2025)	Hong Kong	Quantitative / AES	Custom ChatGPT	Assessment / Feedback	Customised generative AI chatbot for automated essay scoring in disciplinary writing tasks showed stronger rubric alignment than generic ChatGPT prompting.
24	Kohnke (2024)	Hong Kong	Survey / qualitative	ChatGPT / Grammarly	Feedback / EAP	EAP students perceived ChatGPT feedback as more comprehensive and explanatory than Grammarly. Grammarly was preferred for quick surface grammar corrections only.
25	Zhang & Hyland (2023)	Hong Kong	Mixed methods	ChatGPT	Feedback / Engagement	Students engaged more actively with peer feedback than ChatGPT feedback on L2 writing. AI feedback uptake was high but revision quality was comparatively lower.
26	Du & Alm (2024)	New Zealand	Mixed methods	ChatGPT	Feedback / Motivation	Examined ChatGPT impact on EAP students through

						self-determination theory; autonomy-supportive AI interactions positively affected writing motivation.
27	Zhan & Yan (2025)	China	Qualitative	ChatGPT	Feedback literacy	Students' engagement with ChatGPT feedback built feedback literacy progressively. Initial AI reliance diminished as learners became more critical evaluators of feedback.
28	Asadi *et al.,* (2025)	Iran	Quasi-experimental	ChatGPT	Feedback / Writing skills	Integrating ChatGPT with teacher feedback significantly improved Iranian EFL students' writing skills beyond what either source alone could achieve.
29	Mizumoto & Eguchi (2023)	Japan	Quantitative	ChatGPT	AES / Assessment	Explored GPT potential for automated essay scoring. Moderate correlation with human ratings found for linguistic features; recommended as a supplementary scoring tool.
30	Pfau et al. (2023)	USA	Quantitative	ChatGPT	Assessment / Accuracy	Examined ChatGPT for assessing L2 writing accuracy for research purposes. Found moderate reliability; cautioned against using it as the primary scoring mechanism.
31	Wilson & Huang (2024)	USA	Quantitative	Automated system	AES / Fairness	Found evidence of bias in automated essay scoring against elementary-age English language learners; differential performance across language backgrounds documented.
32	Shi & Aryadoust (2024)	Singapore	Systematic review	AWE systems	Assessment / AWE	Systematic review of AI-based automated written feedback research; documented reliability in surface-level scoring and limitations in measuring deep discourse features.
33	Nguyen & Barrot (2024)	Philippines	Empirical / qualitative	ChatGPT	Detection / Assessment	L2 writing teachers struggled to differentiate AI-generated from student-produced

						texts; raises fundamental questions about assessment validity in the AI era.
34	Yancey *et al.* (2023)	USA	Quantitative	GPT-4	AES / CEFR	Rated short L2 essays on the CEFR scale with GPT-4; found adequate alignment with human raters on overall band but divergence on construct-specific criteria.
35	Poole & Coss (2024)	USA	Qualitative / experimental	ChatGPT	AES / Rubric validity	Examined ChatGPT reliability in rubric-based L2 writing assessment; results highly sensitive to prompt design; rubric specificity significantly affected scoring accuracy.
36	Uyar & Büyükahıska (2025)	Turkey	Quantitative	ChatGPT (4o mini)	AES / IELTS	Compared ChatGPT and human rater scores on EFL essays using IELTS band descriptors; significant scoring differences found; human raters showed stronger rubric alignment.
37	Alkhofi (2026)	Saudi Arabia	Quantitative	ChatGPT	AES	Compared ChatGPT AES with human scores on Turkish EFL essays; found weak to moderate correlations; recommended AI as advisory rather than summative scoring tool.
38	Darvin (2025)	Canada	Conceptual / theoretical	ChatGPT / LLMs	Voice / Identity	Theorised voice in L2 writing in the AI era; argued AI-generated text fundamentally lacks authorial positioning and identity expression that defines writer voice.
39	Lucas *et al.* (2023)	Hong Kong	Survey / Conceptual	ChatGPT	Agency / Pedagogy	Examined how ChatGPT reshapes teacher and student roles in L2 writing; argued for explicit AI literacy instruction and redesigned assessment tasks.
40	Ghafouri *et al.* (2024)	Iran	Quasi-experimental	ChatGPT	Self-efficacy / Writing	ChatGPT-based writing instruction protocol improved teacher self-efficacy and learner writing skills over 10 weeks; gains persisted at delayed post-test.

41	Casal & Kessler (2023)	USA	Empirical / corpus	ChatGPT	Detection / Voice	Linguists could not consistently distinguish ChatGPT from human academic writing; raises fundamental questions about authorship, authenticity, and assessment validity.
42	Chen *et al.* (2025)	China	Case study	ChatGPT	Agency / Co-authorship	Followed a Chinese graduate student using ChatGPT throughout academic writing; student considered AI an intellectual collaborator while maintaining primary authorial voice.
43	Strobl *et al.* (2024)	Austria	Action research	ChatGPT	Pedagogy / Writing buddy	Adopting ChatGPT as a writing buddy in advanced L2 writing improved fluency and confidence; structured peer-AI-teacher framework produced the best writing outcomes.
44	Ingley & Pack (2023)	UK	Conceptual	AI tools	Pedagogy / Development	Argued AI tools should develop the writer rather than the writing; proposed task design principles that maintain learner's cognitive engagement with composition process.
45	Wiboolyasarin *et al.* (2024)	Thailand	Quasi-experimental	AI feedback system	Pedagogy / Wiki writing	Synergising collaborative writing and AI feedback in wiki-based environments improved L2 writing proficiency more than either approach independently.
46	Tarchi *et al.* (2024)	Italy	Experimental	ChatGPT	Writing process / Tasks	ChatGPT use in source-based writing tasks reduced cognitive load in synthesis tasks but decreased learners' direct engagement with source material content.
47	Liu *et al.* (2023)	China	Quasi-experimental	AI writing environment	Self-regulation / Writing	AI-supported English writing environments with reflective thinking mechanisms improved both writing quality and metacognitive awareness in EFL learners.

48	Lee *et al.* (2025)	South Korea	Systematic review	GenAI tools	Pedagogy / Language classroom	Systematic review of GenAI in the language classroom; identified formative feedback and vocabulary support as strongest evidence-based applications in L2 contexts.
49	Warschauer *et al.* (2023)	USA	Empirical / mixed	ChatGPT	Equity / Writing affordances	Identified three contradictions in AI-generated text for L2 writers: imitation, the 'rich get richer' effect, and the 'with or without' dilemma regarding AI support.
50	Cotton *et al.* (2024)	UK	Survey / conceptual	ChatGPT	Academic integrity / Equity	Students predominantly did not perceive ChatGPT use as cheating; argued institutions must redefine plagiarism and redesign assessments for the AI era.
51	Alsaedi (2024)	Saudi Arabia	Systematic review	ChatGPT	EFL/ESL writing	Systematic review of ChatGPT in EFL/ESL writing; identified advantages in brainstorming and feedback delivery but limitations in deep-error recognition by AI.
52	Aryadoust & Shi (2023)	Singapore	Systematic review	AWE systems	Assessment / AWE	Systematic review of AWE systems documented gap between surface-level scoring reliability and construct validity for complex discourse features in L2 writing assessment.
53	Li (2025)	Hong Hong	Critical synthesis	GenAI	L2 writing / Review	Synthesised GenAI literacy, written feedback, prompt engineering, and writing assessment validity in one integrated framework for L2 writing research and practice.
54	Tram *et al.* (2024)	Vietnam	Mixed methods	ChatGPT	Writing / Self-learning	ChatGPT served as self-learning English tool among EFL learners; qualitative findings showed learners valued conversational practice and immediate writing corrections.
55	Kofinas *et al.* (2025)	UK	Empirical / conceptual	ChatGPT / GenAI	Academic integrity	Examined GenAI's impact on authentic assessment in higher education; argued for

						redesigning assessments around real-world tasks rather than prohibitive AI policies.
56	Mohamed (2024)	UAE	Survey	ChatGPT	EFL teaching perceptions	EFL faculty members reported positive perceptions of ChatGPT for teaching writing; concerns centred on accuracy of feedback and student over-dependence on AI.
57	Harunasari (2023)	Indonesia	Quasi-experimental	ChatGPT	EFL writing	AI-integrated approach using ChatGPT significantly improved EFL students' writing scores over one semester; strongest gains in grammar and organisation dimensions.
58	Zhang *et al.* (2023)	China	Quasi-experimental	Chatbot	Writing / Argumentation	Chatbot-based learning of logical fallacies in EFL writing improved target knowledge and learner motivation compared with traditional classroom instruction.
59	Jiang *et al.* (2023)	China	Mixed methods	AI writing tools	Assessment / Feedback	Explored AI role in facilitating assessment of L2 writing performance; positive effects on surface accuracy but limited construct coverage found in content assessment.
60	Yeh (2024)	Taiwan	Mixed methods	GenAI tools	Teaching / Inquiry	Synergy of generative AI and inquiry-based learning transformed EFL writing instruction; teachers shifted from knowledge transmitters to critical thinking facilitators.
61	Zhang *et al.* (2025)	China / HK	Comparative (n=60)	GenAI / Hybrid feedback	Feedback / Writing	GenAI feedback improved grammar and sentence variety but limited higher-order skills. Hybrid (GenAI + human) feedback showed greater gains in organisation and critical thinking.
62	Alnemrat *et al.* (2025)	Jordan	Quasi-experimental	ChatGPT	Feedback / EFL writing	AI-generated feedback vs. teacher feedback on EFL argumentative writing at mixed-proficiency levels; AI

						feedback was effective for surface features, teacher feedback for rhetoric.
63	Shi *et al.* (2025)	China	Sequential mixed (n=150)	ChatGPT / AWE	Assessment / Writing self	ChatGPT group scored significantly higher than AWE group after 11 weeks but reported lower ideal L2 writing self; highlights psychological costs of AI feedback use.
64	Deygers *et al.* (2025)	Belgium	Quasi-experimental (n=105)	ChatGPT	AWE / Writing complexity	ChatGPT significantly affected text length and syntactic complexity over 9 weeks but did not produce sustained writing improvement relative to teacher-only feedback.
65	Ding *et al.* (2025)	HK / China	Mixed methods	ChatGPT	AWE / Student perceptions	ChatGPT improved graduate students' writing mechanics, grammar, and overall quality. Combined ChatGPT and instructor feedback led to improvements across all assessed parameters.
66	Hwang *el al.,* (2025)	HK	Screen-recording study	ChatGPT	Writing process (real-time)	L2 learners primarily used ChatGPT for idea generation in early writing stages; later use for vocabulary and polishing was less frequent. Positive perceptions of content creation.
67	Saricaoglu and Bilki (2025)	Turkiye	Empirical / assessment	ChatGPT-4	AES / L2 assessment	Examined ChatGPT-4 capacity for L2 writing assessment for accuracy, specificity, and relevance against analytic rubric; feedback quality varied across writing construct dimensions.
68	Hidayatullah (2024)	Indonesia	Case study	ChatGPT	Writing / Engagement	Investigated ChatGPT for improving EFL writing skills through interactive fun learning activities; students showed improved writing accuracy and higher engagement scores.
69	Nguyen Minh (2024)	Vietnam	Mixed methods	ChatGPT	Writing / Critical thinking	Leveraged ChatGPT for enhancing university freshmen's English writing

						skills and critical thinking; significant gains in argumentation and content development documented.
70	Mali (2025)	Indonesia	Systematic review	ChatGPT	EFL/ESL writing	Systematic review of 32 articles on ChatGPT in EFL/ESL writing classrooms; identified four regulatory categories: institutional policies, instructional strategies, assessment design, ethical co-regulation.
71	Crosthwaite & Sun (2025)	Australia	PRISMA scoping review	GenAI / ChatGPT	Feedback / L2 WF	PRISMA scoping review of 51 SSCI/ESCI studies on GenAI-delivered written feedback in L2 contexts; most studies addressed writing quality improvement and perceptions.
72	Yang *et al.* (2025)	China	Systematic review (n=73)	GenAI / ChatGPT	L2 writing trends	Synthesised 73 empirical studies on GenAI-assisted L2 writing; sociocultural theory most frequently used; ChatGPT dominant; positive and mixed effects with moderate/large effect sizes.
73	Zuo & Zhu (2025)	China	Conceptual	GenAI	Feedback / L2 writing	Synthesised literature on GenAI for L2 written feedback; proposed strategies to optimise GenAI feedback by maximising benefits and minimising weaknesses in feedback provision.
74	Luo *et al.* (2025)	China	Systematic review	AI tools	L2 writing education	Systematic review of 39 empirical studies on AI integration in L2 writing education (2019-2024); documented AWE systems, LLMs, and writing assistants as the three dominant tool types.

2.7 Data Extraction and Synthesis Strategy

Data extraction was conducted using a structured extraction form developed specifically for this review and piloted on ten randomly selected included studies before application to the full sample of 74 studies. The extraction form recorded bibliographic details, country and institutional context, target language, participant characteristics, study design, AI tools used, the focal domain addressed, key findings, equity-relevant dimensions including access barriers and language background effects, and voice or agency-relevant dimensions including authorship discussions

and learner autonomy findings Ouzzani *et al.,* (2016). Extraction was completed independently by two reviewers for a random 25 percent sub-sample to verify consistency; agreement on factual fields was 94 percent, and differences were resolved by returning to the source text. Interpretive coding fields were discussed between reviewers for the full sample to ensure consistent application of coding categories.

The synthesis followed a framework synthesis approach, in which an analytical framework derived from the review questions and prior theory is applied to the extracted data and refined iteratively as patterns emerge Page *et al.*, (2013). The four conceptual tensions identified in the introduction served as the initial analytical framework. As coding progressed across the 74 included studies, sub-themes within each tension were identified and recorded in an evolving thematic map. Quantitative findings were compared narratively rather than statistically because the heterogeneity of outcome measures, study designs, and population characteristics across the quantitative studies did not meet conditions required for meta-analytic pooling Page *et. al.,* (2021).

3. Publication Trends and Research Distribution

The volume of research on generative AI in L2 writing has grown faster than perhaps any comparable research area in applied linguistics. Before 2023, the field was dominated by studies of automated writing evaluation and automated essay scoring systems, with relatively little attention to generative tools Moorhouse *et al.,* (2025). The release of ChatGPT in November 2022 triggered an immediate and sustained surge of empirical and conceptual work that has not slowed as of the final search date. Studies indexed in 2023 and 2024 account for the substantial majority of the included literature. This publication rate creates its own methodological problem: peer review timelines in many journals have been compressed by editorial pressure to address urgent questions quickly, and this pressure may have affected the rigour of some of the work that passed through review.

Geographically, the research is heavily concentrated in East and Southeast Asia, with China, Taiwan, South Korea, and Thailand contributing the largest numbers of empirical studies. This geographic concentration reflects both the large EFL student populations in these regions and the institutional pressure on scholars in these settings to publish in high-impact international journals. Studies from Sub-Saharan Africa, South Asia, Latin America, and the Middle East are not well represented. This is a pattern with serious implications for generalising findings about AI writing tools whose primary design language is English. Studies from Europe tend to focus on higher education contexts, while studies from Asia address a wider range of educational levels including secondary school and language institutes. The under-representation of primary and secondary educational contexts in the research is a notable gap, given that AI writing tools are increasingly used by younger learners whose language development trajectories differ substantially from those of university students (Table 5).

Table 5. Summary of publication trends in the reviewed literature (2022–2026)

Characteristic	Distribution in Reviewed Literature
Publication years	2022: sparse; 2023: major growth; 2024–2026: sustained high volume
Geographic focus	East/Southeast Asia (dominant); Europe; North America; others under-represented
Educational level	Higher education (majority); secondary school; language institutes
Target language	English as L2 (dominant); Chinese, Arabic, French, Spanish (limited)
Paper type	Empirical quantitative; empirical qualitative; mixed methods; systematic reviews; conceptual
AI tools studied	ChatGPT/GPT-4 (most frequent); Grammarly; WriteAhead; e-rater; proprietary tools

In terms of target languages, English as a foreign or second language dominates the literature so completely that it is difficult to draw firm conclusions about generative AI in writing contexts for any other target language Hakim, (2023). There is a small but growing body of work on AI tools for Chinese, Arabic, French, and Spanish as foreign languages, but these studies rarely appear in the synthetic reviews that shape the field's direction. The dominance of English in both the research literature and the design of AI writing tools creates a feedback loop: AI tools perform better for English than for other languages, studies find more positive effects for English contexts, and

investment and development continue to favour English. This loop has serious equity implications that are addressed in Section 8.

4. Generative AI Across the Writing Process

4.1 Brainstorming and Idea Generation

Research on generative AI in the brainstorming phase consistently reports that learners find AI tools useful for overcoming the blank page problem, generating topic ideas, identifying relevant angles for argumentation, and organising initial thoughts Cheng *et al.,* (2023). Studies conducted in EFL university contexts in China and South Korea find that students who use AI tools for brainstorming report lower writing anxiety and higher confidence in beginning the writing task Lee *et al.,* (2024). These findings are consistent with what the broader second language writing literature has established about the role of pre-writing support in reducing the cognitive load of L2 composition. However, a recurring concern in qualitative studies is that AI generated brainstorming ideas may substitute for the learner's own cognitive engagement with the topic rather than scaffolding it. When a learner asks ChatGPT to generate five arguments for a position and then writes from those arguments without engaging in their own elaboration, the brainstorming phase effectively bypasses the generative thinking that is itself a site of language learning Hakim, (2023).

4.2 Outlining and Planning

The role of AI tools in the outlining phase has received less systematic attention than brainstorming or drafting, but available evidence suggests that learners use AI assisted outlining in two distinct ways. Some learners use AI to generate a complete structural plan for an essay, which they then fill in with their own content. Others use AI as a check on an outline they have already drafted, asking the AI to identify gaps, logical inconsistencies, or missing counterarguments Westbye *et al.,* (2026). The second use pattern is more consistent with process-oriented writing pedagogy, which emphasises planning as a metacognitive activity that develops learners' awareness of text structure and argumentation. The first use pattern may reduce the planning phase to an act of retrieval and execution rather than strategic thinking. Teachers who design AI integrated writing tasks need to be explicit about which use of AI during the planning phase they are supporting, because the two patterns have different implications for what learners develop through the writing task.

4.3 Drafting

Drafting with AI assistance is the most contested phase in the current literature. At one end of the spectrum, studies document learners who use AI tools primarily for local support during drafting, asking for vocabulary suggestions, sentence reformulations, or explanations of grammatical structures as they compose Sahan *et al.,* (2023). At the other end, studies document learners who submit a prompt and substantially copy the AI generated response, engaging minimally with their own composition. Both patterns appear in the same institutional contexts, and neither is easily explained by learner proficiency level or motivation alone. The research suggests that the distinction between AI supported drafting and AI substituted drafting is determined less by learner characteristics than by task design. When writing tasks require personalised content, local knowledge, or explicitly positioned argument, learners are less likely to submit unmodified AI text because such text does not address the task adequately Ferber *et al.,* (2025) This finding has a practical implication: teachers can reduce wholesale AI substitution by designing tasks that require the learner's specific perspective, experience, or cultural knowledge. Generic argument tasks on abstract topics are precisely the ones most vulnerable to AI replacement.

4.4 Revision and Editing

The revision phase is where the current literature is most positive about the role of AI tools in L2 writing development. Studies consistently find that AI generated feedback on drafts leads to measurable improvements in subsequent versions of student texts Wiboolyasarin*, et al.,* (2024) Learners report that AI feedback is available immediately, responds to specific aspects of their text, and explains the reasons for suggestions in accessible

language. These properties address known weaknesses of peer feedback in L2 writing contexts, where feedback quality is often limited by the peers' own proficiency level. The relationship between acting on AI feedback and developing as a writer is more complicated, however. Studies that look beyond immediate text improvement find that learners who rely heavily on AI for revision develop less critical awareness of their own recurring errors than learners who engage with teacher feedback that asks them to identify patterns across multiple texts Hakim, (2023). The editing phase — which involves final surface level correction — is where AI tools such as Grammarly and ChatGPT based checkers show the most consistent and least contested positive effects. Even here, though, questions arise about whether correction without explanation supports accuracy development over the long term.

5. Generative AI for Written Feedback

5.1 Corrective Feedback

Corrective feedback is the most extensively studied dimension of AI feedback in L2 writing, and research in this area stretches back to the earliest days of automated writing evaluation (Lo *et al.,* (2025) Generative AI has expanded the scope and sophistication of what automated systems can provide. Where earlier tools were limited to detecting surface errors in grammar, spelling, and punctuation, large language model based systems can now identify cohesion failures, argumentative gaps, register inconsistencies, and violations of discourse conventions. Studies that compare AI corrective feedback with teacher corrective feedback find that AI feedback is more consistent across learners and more thorough in identifying surface errors, while teacher feedback tends to be more attuned to the specific communicative intent of the writer and the developmental priorities appropriate to that learner Aryadoust *et al.,* (2020).

5.2 Formative Feedback

Formative feedback in writing education refers to feedback aimed at supporting learning and development rather than simply evaluating a finished product. The most promising applications of generative AI in formative feedback involve the use of AI tools to provide ongoing, responsive feedback throughout the writing process rather than at a single summative point Li & Collins (2026). Studies in this area find that learners who receive AI formative feedback at multiple stages of a writing task show greater improvement across drafts than learners who receive a single round of feedback at the end. This finding is consistent with what the process-oriented writing literature has established about the timing of feedback. However, a critical question that few studies address is whether AI formative feedback supports the development of self-regulatory writing skills. Learners who receive immediate, detailed AI feedback at every stage may improve their immediate text quality while becoming dependent on AI support rather than developing the metacognitive awareness to self-monitor and self-correct in contexts where AI is not available Alsaiari *et al.,* (2026).

5.3 Dialogic Feedback and Teacher-AI Combinations

Dialogic feedback, in which learner and feedback provider engage in iterative exchange to negotiate the meaning and application of feedback suggestions, represents an important theoretical framework for understanding how feedback supports writing development Steen-Utheim, & Wittek, (2017). Generative AI tools, particularly conversational systems such as ChatGPT, have dialogic affordances that earlier automated writing evaluation tools lacked entirely. A learner can ask a generative AI to explain why it made a particular suggestion, to offer alternatives, or to clarify how a suggested change improves the text. Studies that have examined learner interaction logs in AI writing feedback contexts find that learners who engage dialogically with AI feedback, asking follow-up questions and requesting elaborations, show greater revision quality than learners who simply accept or reject suggestions without further engagement Escalante *et al.,* (2023) Teacher and AI combinations represent a promising middle ground in this area. AI handles high-volume, consistent feedback on surface features while teachers focus their limited time on higher-order concerns about meaning, argument, and writer development. This division of labour is not simply a practical convenience. It reflects a principled understanding of what each feedback source does best. Several studies have found that combined AI and teacher feedback produces better writing outcomes than either source alone Asadi *et al.,* (2025) which suggests that the two sources are complementary rather than competing.

6. Generative AI in Writing Assessment

6.1 Scoring and Rubric Alignment

The use of large language model based systems for essay scoring represents a significant departure from earlier automated scoring approaches. Classical automated essay scoring systems relied on features extracted from text i.e., lexical diversity, sentence length variation, syntactic complexity that served as proxies for writing quality Atkinson & Palma (2025). Contemporary AI scoring systems use models trained on large human-scored corpora and are capable of aligning their scores with specific rubric criteria in ways that earlier systems could not. Studies that have evaluated the rubric alignment of large language model based scoring find moderate to high correlations with expert human scoring on holistic scales and dimension-specific rubrics in EFL contexts (Aryadoust, *et al.*, 2020). These correlation figures mask important patterns of divergence. AI scores tend to align most closely with human scores on surface accuracy and organisation, and least closely on content quality and writer voice. Content quality and writer voice are precisely the constructs that L2 writing educators consider most important for developmental assessment purposes. A scoring system that is highly reliable on dimensions that educators consider less important, and less reliable on dimensions they consider most important, is not a valid assessment tool even if its average correlation with human scores appears respectable.

6.2 Reliability, Validity, and Fairness

The reliability of AI scoring systems in L2 writing contexts is, in narrow technical terms, high. Automated systems score the same essay the same way every time, a property that human raters, who are subject to fatigue, context effects, and inter-rater variability, cannot match Wang *et al.,* (2025). This surface reliability is sometimes presented as an unambiguous advantage of AI scoring, but it conflates consistency with validity. A system that consistently scores a poorly measured construct is consistently invalid. Validity in writing assessment requires not just that scores are consistent but that they represent the construct the assessment is intended to measure, that they support appropriate interpretations and uses, and that their consequences are justifiable (Aryadous *et al.*, 2020).

Current evidence on the validity of AI scoring in L2 contexts raises several serious concerns. The training data for most commercial AI scoring systems consists predominantly of first language writing, meaning that what the system has learned to recognise as high quality writing reflects first language discourse norms. AI scoring systems have also been shown to reward certain surface features of academic writing that can be produced without corresponding depth of thought, a problem sometimes called the gaming effect. The opacity of most commercial AI scoring systems makes it impossible to verify what the system is actually measuring, which is a fundamental obstacle to any serious validity argument. Fairness in AI writing assessment has emerged as a particularly important concern as evidence accumulates about differential system performance across learner groups. Systems trained primarily on English first language corpora tend to penalise the discourse features characteristic of writing by speakers of Arabic, Chinese, and Japanese, even when those features reflect legitimate rhetorical conventions rather than errors Aydın *et al.,* (2025). Differential item functioning analyses of AI scoring systems in multilingual assessment contexts consistently find significant performance gaps that cannot be explained by proficiency differences alone. These fairness concerns have direct consequences for the use of AI scoring in high-stakes contexts such as university admissions, scholarship decisions, and English language proficiency certification.

6.3 Transparency and Accountability

The transparency of AI writing assessment systems is a serious problem that the current literature has not adequately addressed. Most commercial AI scoring tools are proprietary, and their developers are not required to disclose the training data, scoring algorithms, or validation evidence that would allow independent evaluation of their claims BaHammam (2025). This opacity creates a fundamental accountability gap in high-stakes assessment contexts. Institutions that adopt commercial AI scoring systems are, in effect, delegating consequential assessment decisions to systems whose internal logic they cannot inspect or contest. The research literature has called for minimum transparency standards for AI assessment tools used in educational contexts, including the disclosure of training data demographics, the publication of differential performance analyses across learner groups, and the availability of human review mechanisms for consequential decisions. These calls have not yet been reflected in consistent regulatory or policy action.

7. Writer Voice, Agency, Identity and Authorship

Writer voice is one of the most theoretically complex and pedagogically significant constructs in second language writing research. It encompasses a writer's distinctive patterns of linguistic choice, their stance toward their topic and audience, their expression of personal or cultural identity through textual means, and the sense of ownership and intentionality that characterises engaged writing Mhilli, (2023). For L2 writers, voice development is particularly important because it is the dimension of writing in which linguistic development and identity development intersect most directly. Writing in a second language is not simply a technical skill. It is an act of self-presentation in a new linguistic and cultural medium, and the choices a writer makes about vocabulary, stance, structure, and rhetorical strategy reflect and shape their identity as a member of the target language community. The arrival of generative AI raises urgent questions about what happens to voice when the text a learner submits has been substantially generated or modified by a system that has no identity, no stake, and no cultural positioning.

The available research on voice and AI writing presents a clear divided representation. On one side, studies document L2 learners who report that AI tools help them express their ideas more effectively than they could in their own interlanguage, effectively extending their communicative reach beyond what their current proficiency would allow Wang & Wang, (2025); Chen (2025). On the other side, qualitative studies that examine learner texts before and after AI revision find that AI generated modifications tend to produce more uniform, conventionally academic texts that are harder to distinguish from one another and show fewer traces of the writer's individual linguistic choices Chen *et al* (2026); Fariello, *et al.,* (2025). This finding is consistent with what is known about the training data of large language models, which is dominated by high-frequency, conventionally successful academic writing and therefore tends to produce text that conforms to the dominant register rather than exploring its margins. The result is a kind of averaging effect: AI revision makes texts more similar to one another and to an implicit standard, at the cost of the individual variation that writers develop through taking linguistic risks.

The question of authorship has become one of the most contested issues in academic writing in the wake of generative AI. Traditional conceptions of authorship assume a human author who is responsible for the ideas, the argument, and the linguistic expression in a text Amirjalili *et al.,* (2024). When a significant portion of a text is generated by an AI system, this conception becomes difficult to sustain, yet it underlies the assumptions of academic integrity policies, assessment rubrics, and the very purpose of writing tasks in educational contexts. The research literature has not resolved this question. Some scholars argue that AI use in writing is no different from the use of a dictionary or a grammar checker: it is a tool the writer uses, and the writer remains the author. Others argue that generative AI's capacity to produce extended, coherent, original argument is qualitatively different from earlier tools and that its use in assessed writing constitutes a fundamental distortion of what assessment is designed to measure.

8. Equity, Multilingual Access and Digital Divides

The equity dimensions of generative AI in L2 writing are among the most important and least studied aspects of the current landscape. The dominant narrative in technology adoption research is one of democratisation: new tools lower barriers to access and allow disadvantaged groups to benefit from capabilities previously available only to the privileged Kim *et al.*, (2025). This narrative is not wholly wrong in the AI writing context. A learner in a rural language institute with no access to a trained writing tutor can ask ChatGPT for feedback on their draft and receive a detailed, grammatically sophisticated response. This is a genuine form of access extension. The democratisation narrative, however, obscures the ways in which AI writing tools reproduce and amplify existing inequalities along several dimensions simultaneously.

The first dimension of inequity is linguistic: AI writing tools perform substantially better for English than for other languages, and within English, they perform better for the standard academic variety of English associated with educated first language speakers than for the varieties of English produced by speakers of different first languages Prakash *et al.,* (2025). A learner whose first language is Arabic or Chinese faces a double disadvantage: the AI tool they use is trained primarily on English language data and therefore provides less contextually appropriate feedback for texts that reflect Arabic or Chinese discourse conventions, and the feedback the tool provides tends to push their writing toward a standard English norm that may not reflect the communicative context for which they are actually writing. The second dimension is economic: premium AI writing tools require subscription fees that are substantial relative to average household incomes in many of the countries where large EFL populations exist. Free

tiers of these tools have limited functionality and usage caps that the most sophisticated AI writing support is effectively available only to learners who can afford to pay for it.

The third dimension of inequity is infrastructural. Reliable, high-speed internet connectivity is required to use cloud-based AI writing tools effectively, and large portions of the world's EFL student population study in environments where such connectivity cannot be assumed Yu *et al.*, (2026). Studies of AI tool adoption in Sub-Saharan Africa, South Asia, and parts of Southeast Asia consistently identify connectivity and device access as the primary barriers to use, barriers that are not addressed by the development of better AI models. The digital divide in AI writing tool access therefore reproduces and potentially amplifies the educational inequalities that existed before generative AI, because learners in under-resourced contexts are denied the efficiency and feedback benefits that their better-resourced counterparts enjoy. Research that evaluates AI writing tools in high-resource institutional contexts and then generalises to global EFL education systematically overstates the equity benefits of AI adoption.

9. Pedagogical Design and Teacher Roles

Pedagogical design is the dimension of the current literature that shows the greatest gap between what is known and what is needed. The research is rich in studies that evaluate the effects of AI tools on writing quality, feedback uptake, and assessment outcomes, but it is relatively thin in studies that examine what makes an AI integrated writing pedagogy coherent, principled, and effective over time Tsao *et al.,* (2025). The distinction between using an AI tool in a writing class and designing an AI integrated writing pedagogy is important (Figure 3). Any teacher can allow students to use ChatGPT for feedback on their drafts. A much smaller number of teachers have thought through what role AI feedback should play in relation to teacher feedback, peer feedback, and self-assessment; what writing tasks are well suited to AI assistance and which are better done without it; how to scaffold learners' critical engagement with AI suggestions rather than passive acceptance; and how to assess writing development in ways that are meaningful in an AI-rich environment.

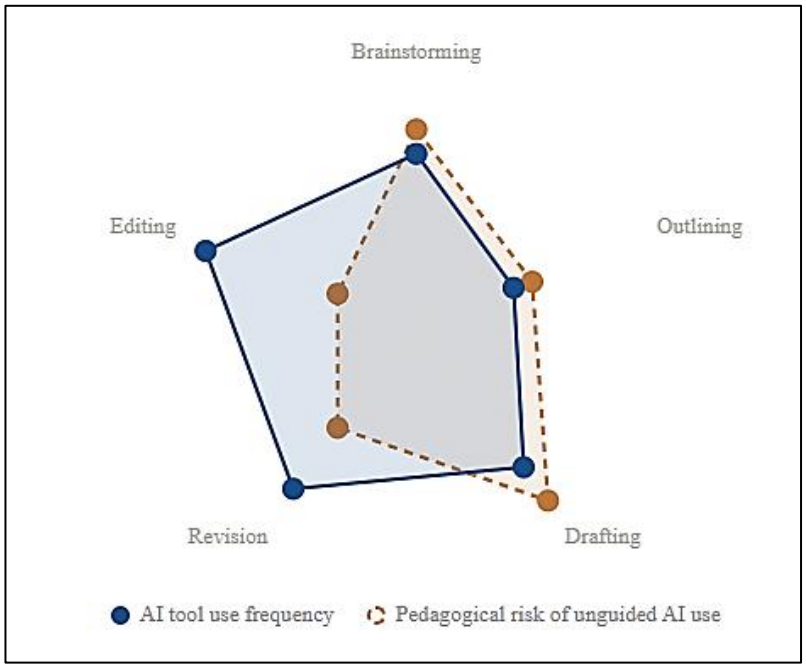

Figure 3. AI use and pedagogical risk by writing phase

The teacher's role in an AI integrated writing classroom is being reconceptualised in the current literature, though the reconceptualisation is not yet complete or consistent Chaieb *et al.,* (2026). Some studies describe the teacher as a curator of AI tools, selecting appropriate tools for different tasks and teaching learners to use them critically. Others describe the teacher as a metacognitive coach, whose primary role shifts from providing content feedback to helping learners understand their own writing development in relation to AI-mediated feedback. Still others argue that the teacher's most important role in an AI-rich environment is ethical guide, helping learners navigate the complex questions of authorship, integrity, and identity that AI use raises. These roles are not mutually exclusive. The most thoughtful accounts of AI integrated writing pedagogy combine them. Pre-service and in-service

teacher education has not yet developed the frameworks, materials, or assessment tools to prepare teachers for these expanded and reconfigured roles Seufert *et al.,* (2025).

Task design is a critical but not well-researched dimension of AI integrated writing pedagogy. The available evidence suggests that the nature of the writing task is the most powerful determinant of how learners use AI tools (Hakim, 2023). Tasks that require learners to draw on personal experience, local knowledge, or explicitly positioned argument are more resistant to wholesale AI substitution than tasks that ask for generic argumentation on abstract topics. Tasks that require learners to explain their AI use, reflect on AI feedback, or compare AI generated text with their own drafts build critical AI literacy alongside writing skills. Tasks that include process portfolios, revision histories, and reflection journals create accountability for the writing process as well as the product. These design principles are consistent with what the broader writing pedagogy literature has established about task design for process-oriented writing instruction.

10. Methodological Weaknesses in the Current Literature

The current literature on generative AI in L2 writing is characterised by several patterns of methodological weakness that limit the confidence with which findings can be interpreted and the degree to which conclusions can be generalised. The first and most pervasive weakness is small sample size. The majority of empirical studies in the literature reviewed involved fewer than 60 participants, and many involved fewer than 30 Effatpanah *et al.,* (2026). Small samples create serious problems for statistical power in quantitative studies and limit the transferability of qualitative findings. The small sample sizes in this literature reflect the genuine difficulty of conducting well-controlled studies of AI writing tool use in authentic educational settings. The cumulative effect, however, is a body of evidence that is fragmented, inconsistent, and difficult to synthesise.

The second major methodological weakness is the absence of longitudinal data. The research questions that matter most for L2 writing education are not about whether AI tools improve a specific text but whether AI integrated writing instruction produces better writers over time Niekerk *et al.,* (2025) Answering this question requires studies that follow learners across multiple writing tasks, multiple feedback cycles, and multiple semesters, and compare their trajectories with those of learners who received different types of instruction. Such studies are rare in the current literature. Most studies examine a single writing task or a single course, with pre and post measurements that capture only a very short developmental window.

The third weakness is the absence of strong theoretical grounding for the mechanisms through which AI tools affect writing development. Many studies demonstrate that AI feedback leads to improved revision or that AI assisted writing produces higher-scoring texts, but relatively few studies investigate why or how these effects occur Lo *et al.,* (2025). Theoretical frameworks from second language acquisition, writing studies, and educational psychology such as sociocultural theory, the noticing hypothesis, self-regulation theory, and dynamic systems theory are referenced in many papers but rarely used as the basis for specific hypotheses that the study then tests. The result is a literature that accumulates findings without accumulating explanation.

The fourth weakness is over-reliance on self-reported attitudinal data. Many studies measure learner satisfaction with AI tools, perceived usefulness, and self-reported improvement without examining actual writing quality or without triangulating self-report data with text analysis or interaction logs. Self-reported perceptions of AI usefulness are valuable as data, but they should not substitute for analysis of writing outcomes. A learner who reports that ChatGPT helped them write better is not necessarily a learner who has actually developed as a writer. The field needs to invest in outcome measures that can distinguish between AI mediated text improvement and genuine writer development.

11. A Research Agenda: Validity, Voice, Equity, and Human-AI Co-Authorship

The most urgent research priority in this field is establishing the construct validity of AI scoring systems for L2 writing in contexts and for learner populations that differ from the training data on which those systems were built Ren *et al.,* (2025). This requires systematic studies that examine not just correlations between AI scores and human expert scores but the construct representations that AI systems use, the population invariance of scoring across language backgrounds, and the consequential validity of AI based assessment decisions in educational

settings. Validity studies of this kind require collaboration between applied linguists, measurement scholars, and AI scoring system developers. They also require access to training data and system architectures that are currently unavailable for proprietary systems. Advocacy for transparency in AI assessment systems is therefore not just an ethical position but a prerequisite for the validity research the field needs.

The second research priority is a sustained investigation of voice, authorship, and identity in human and AI collaborative writing. This research needs to move beyond the binary question of whether AI use constitutes academic dishonesty and address the more fundamental question of what it means to develop as a writer in a world where AI writing support is widely available Hakim (2023). Studies that track the same learners' writing development across contexts with and without AI support, using both text analysis and qualitative accounts of learner experience, would contribute significantly to understanding the developmental effects of AI writing tool use. Research is also needed on how different types and levels of AI engagement from local vocabulary assistance to wholesale draft generation differentially affect the development of writer voice, self-regulatory skills, and metalinguistic awareness.

The third research priority is equity focused investigation of AI writing tool access and effects across diverse learner populations and institutional contexts Marzuki *et al.,* (2023). This research should not be limited to documenting access gaps. It should include studies that design and evaluate interventions specifically intended to extend access and adapt AI tools for populations and languages that are not well served at present. The fourth priority is developing and evaluating principled pedagogical frameworks for AI integrated writing instruction. The research needs to move from documenting what AI tools can do to establishing what teachers and institutions should do with them to support writing development, maintain academic integrity, and advance equity.

12. Implications for Teachers, Institutions, and Journal Editors

For teachers of L2 writing, the most actionable implication of this review is that AI integrated writing pedagogy requires explicit design rather than default accommodation. Allowing students to use AI tools without designing specific tasks, scaffolds, and reflection activities that direct how AI is used is not AI integrated pedagogy. It is AI-permissive instruction whose effects on writing development are likely to be inconsistent and unpredictable (Ramadani *et al.,* 2025). Teachers who want to use AI tools to support writing development should begin by identifying specific tasks in their curriculum where AI assistance is likely to extend learner capability in ways that are consistent with course learning outcomes, and distinguish those clearly from tasks where unassisted writing is the goal. They should design tasks that require learners to engage critically with AI feedback rather than simply accept or reject it, and they should make space in their courses for explicit discussion of the ethical, identity, and authorship questions that AI use raises.

For institutions, the implications concern both assessment policy and professional development. Assessment policies on AI use need to move beyond blanket prohibition or blanket permission toward differentiated guidance that specifies, for different assessment purposes and contexts, what AI assistance is acceptable and what constitutes unacceptable substitution Lyu *et al.,* (2025) Institutions that use commercial AI scoring tools in high-stakes contexts should require the same disclosure, validation, and differential performance analysis from AI assessment providers that they would require from human raters. They should invest in professional development that gives teachers the theoretical understanding and practical skills to design AI integrated writing instruction, and they should provide access to AI writing tools in ways that do not create additional inequities between students who can afford premium subscriptions and those who cannot.

For journal editors and reviewers, the implication is that submissions on AI in L2 writing should be held to the same methodological standards as other empirical work. This means requiring adequate sample sizes, longitudinal designs where the research question demands them, and theoretical frameworks that are not merely decorative. Editors should be especially attentive to the gap between claims about AI's effects on writing development and the actual evidence base that any given study can provide. Short-term studies that measure text quality on a single task cannot support conclusions about writer development, and reviewers should push back when authors make that inferential leap.

13. Conclusion

This review addressed that literature on generative AI in L2 writing, feedback, and assessment has grown faster than the methodological tools needed to evaluate it. The need for such a review arose from the convergence of three conditions: the speed and scale of AI adoption in writing-relevant educational settings, the fragmentation of the research across disciplinary silos, and the absence of a synthetic, critically informed account of what is known, what is uncertain, and what needs to be done. The review drew on PRISMA 2020 methodology to identify and evaluate 74 studies meeting quality appraisal thresholds, synthesised findings across the three domains, and identified four cross-cutting tensions as the organising themes of the literature. The first contribution of this review is a comprehensive synthesis that integrates the three domains of writing support, feedback, and assessment within a unified analytical frame. This integration revealed that the four tensions i.e, validity versus efficiency, voice versus support, equity versus access, and capability versus design, run through all three domains simultaneously and cannot be resolved within any single domain in isolation. The second contribution is a sustained methodological critique that identified small sample sizes, absence of longitudinal designs, accounts that are not well grounded in theory, and over-reliance on self-report as systemic weaknesses that constrain the conclusions the current literature can support. The third contribution is a research plan organized around four priorities: equitable and multilingual access, voice and authorship in AI-assisted writing, construct validity in AI grading, and principled pedagogical design. These priorities are both theoretically grounded and practically significant.

The most important direction for future research is longitudinal investigation of writing development in AI rich environments that asks not whether AI tools improve a specific text but whether and how they shape the development of writers over time. This research should attend to the full range of L2 contexts and learner populations, not only the high-resource university contexts that currently dominate the literature. It should be built on explicit theoretical frameworks that connect AI tool use to established accounts of second language writing development and that generate testable predictions rather than post hoc explanations. The field of L2 writing has always had to negotiate the tension between technological change and pedagogical purpose. Generative AI has intensified this tension to a degree that makes principled, evidence-based scholarship more important than ever. The purpose of such scholarship is not to determine whether AI belongs in writing education but to ensure that it serves the writers whose development is, and must remain, the central concern.

References

Alkhofi, A. (2026). Can ChatGPT score ESL writing? A Correlation Analysis between Teacher and GenAI scores. *Language Teaching Research*. https://doi.org/10.1177/13621688261415584

Alnemrat, A., Aldamen, H., Almashour, M., Al-Deaibes, M., AlSharefeen, R. (2025). AI vs. Teacher Feedback on EFL Argumentative Writing: a Quantitative Study. *Frontiers in Education*, *10*. https://doi.org/10.3389/feduc.2025.1614673

Alsaedi, N. (2024). ChatGPT and EFL/ESL Writing: A Systematic Review of Advantages and Challenges. *English Language Teaching*, 17(5), 41. https://doi.org/10.5539/elt.v17n5p41

Alsaiari, O., Baghaei, N., Lodge, J.M., Noroozi, O., Gašević, D., Boden, M., Khosravi, H. (2026). Directive, Metacognitive, or a Blend of Both? A comparison of AI-Generated Feedback types on Student Engagement, confidence, and outcomes. Computers and Education: Artificial Intelligence, 100553. https://doi.org/10.1016/j.caeai.2026.100553

Amirjalili, F., Neysani, M., Nikbakht, A. (2024, March). Exploring the Boundaries of Authorship: A Comparative Analysis of AI-generated text and human Academic Writing in English Literature. In Frontiers in Education Frontiers Media, 9, 1347421. https://doi.org/10.3389/feduc.2024.1347421

Aryadoust, V., Ng, L.Y., Sayama, H. (2020). A Comprehensive Review of Rasch Measurement in language Assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Asadi, M., Ebadi, S., Mohammadi, L. (2025). The Impact of Integrating ChatGPT with Teachers' Feedback on EFL Writing Skills. *Thinking Skills and Creativity*, *56*, 101766. https://doi.org/10.1016/j.tsc.2025.101766

Atkinson, J., Palma, D. (2025). An LLM-based hybrid approach for enhanced automated essay scoring. Scientific Reports, 15(1), 14551. https://doi.org/10.1038/s41598-025-87862-3

Aydın, B., Kışla, T., Elmas, N.T., Bulut, O. (2025). Automated Scoring in the Era of Artificial Intelligence: An Empirical Study with Turkish essays. System, 133, 103784. https://doi.org/10.1016/j.system.2025.103784

BaHammam, A.S. (2025). The Transparency Paradox: why Researchers Avoid Disclosing AI Assistance in Scientific Writing. Nature and Science of Sleep, 2569-2574. https://doi.org/10.2147/NSS.S568375

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

Bender, E.M., Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bour, C., Schmitz, S., Ahne, A., Perchoux, C., Dessenne, C., Fagherazzi, G. (2021). Scoping review Protocol on the use of Social Media for Health Research Purposes. *BMJ Open*, *11*(2), e040671. https://doi.org/10.1136/bmjopen-2020-040671

Casal, J.E., Kessler, M. (2023). Can Linguists Distinguish between ChatGPT/AI and Human Writing?: A study of Research Ethics and Academic Publishing. *Research Methods in Applied Linguistics*, *2*(3), 100068. https://doi.org/10.1016/j.rmal.2023.100068

Chaieb, M., Cuel, R., Bouzaabia, R. (2026). Reconceptualizing of Genai Adoption in Higher Education: A Task-Based Perspective. The International Journal of Management Education, 24(2), 101365. https://doi.org/10.1016/j.ijme.2026.101365

Chen, C., Gong, Y. (2025). The role of AI-Assisted Learning in Academic Writing: A Mixed-Methods Study on Chinese as a Second Language students. *Education Sciences*, *15*(2), 141. https://doi.org/10.3390/educsci15020141

Chen, S.Y.R. (2026). Two voices, one draft: a Diachronic Comparison of Human and AI feedback in EFL Essay Writing. Australian Journal of Applied Linguistics, 9, 103200-103200. https://doi.org/10.29140/ajal.2026.103200

Chen, Y. (2025). Evaluating the Potential of ChatGPT-Reformulated Essays as Written Feedback in L2 Writing. Computers and Education: Artificial Intelligence, 100500. https://doi.org/10.1016/j.caeai.2025.100500

Cheng, Z., Wang, H., Zhu, X., West, R.E., Zhang, Z., Xu, Q. (2023). Open Badges Support Goal Setting and Self-Efficacy but not Self-Regulation in a hybrid-learning environment. Computers & Education, 197, 104744. https://doi.org/10.1016/j.compedu.2023.104744

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Conijn, R., Martinez-Maldonado, R., Knight, S., Shum, S.B., Van Waes, L., Van Zaanen, M. (2020). How to provide Automated Feedback on the Writing Process? A Participatory Approach to Design Writing Analytics tools. *Computer Assisted Language Learning*, *35*(8), 1838–1868. https://doi.org/10.1080/09588221.2020.1839503

Cotton, D.R.E., Cotton, P.A., Shipway, J.R. (2023). Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Crosthwaite, P., Sun, S. (2025). Generative AI and L2 Written Feedback Studies: A scoping review. *RELC Journal*, 57(1), 207–219. https://doi.org/10.1177/00336882251386530

Darvin, R. (2025). The need for Critical Digital Literacies in Generative AI-Mediated L2 Writing. *Journal of Second Language Writing*, *67*, 101186. https://doi.org/10.1016/j.jslw.2025.101186

Deygers, B., Buelens, L., Chan, D., Schildt, L., Van Parys, A., Vanbuel, M. (2025). The Impact of Automated Writing Evaluation on Writing Gains. *ELT Journal*, *79*(3), 352–362. https://doi.org/10.1093/elt/ccaf020

Ding, L., Zou, D., Kohnke, L. (2025). ChatGPT as an Automated Writing Evaluation tool: how students Perceive it and how it affects their Writing. *Education and Information Technologies*, *30*(18), 26777–26799. https://doi.org/10.1007/s10639-025-13775-3

Du, J., Alm, A. (2024). The Impact of CHATGPT on English for Academic Purposes (EAP) Students' Language Learning Experience: A Self-Determination Theory Perspective. *Education Sciences*, *14*(7), 726. https://doi.org/10.3390/educsci14070726

Effatpanah, F., Ravand, H., Tabari, M.A., Chen, Y.H., Kunina-Habenicht, O. (2026). How do L2 writing Subskills Interact Hierarchically? Insights from diagnostic classification Models. Assessing Writing, 68, 101029. https://doi.org/10.1016/j.asw.2026.101029

Efthimiou, O., Debray, T.P.A., Van Valkenhoef, G., Trelle, S., Panayidou, K., Moons, K.G.M., Reitsma, J. B., Shang, A., Salanti, G., Group, O.B.O.G.M.R. (2016). GetReal in Network Meta-Analysis: a Review of the Methodology. *Research Synthesis Methods*, *7*(3), 236–263. https://doi.org/10.1002/jrsm.1195

Escalante, J., Pack, A., Barrett, A. (2023). AI-generated feedback on writing: insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20(1). https://doi.org/10.1186/s41239-023-00425-2

Fariello, S., Fenza, G., Forte, F., Gallo, M., Marotta, M. (2025). Distinguishing Human from Machine: A review of Advances and Challenges in Ai-Generated Text Detection. International Journal of Interactive Multimedia and Artificial Intelligence, 9(3), 6-18. https://doi.org/10.9781/ijimai.2024.12.002

Ghafouri, M., Hassaskhah, J., Mahdavi-Zafarghandi, A. (2024). From Virtual Assistant to Writing Mentor: Exploring the Impact of a ChatGPT-based Writing Instruction Protocol on EFL Teachers' Self-Efficacy and Learners' Writing Skill. *Language Teaching Research*. https://doi.org/10.1177/13621688241239764

Godwin-Jones, R. (2024). Distributed agency in second language learning and teaching through generative AI. *Language Learning & Technology*, 28(2), 5–31. https://doi.org/10.64152/10125/73570

Guo, K., Wang, D. (2023). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, 29(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Hakim, A. (2023). Genre-Related Episodes as a Lens on Students' Emerging Genre Knowledge: Implications for Genre-Based Writing Pedagogy, Collaborative Tasks, and Learning Materials. *Journal of Second Language Writing*, *60*, 101001. https://doi.org/10.1016/j.jslw.2023.101001

Harunasari, S. Y. (2023). Examining the Effectiveness of AI-Integrated Approach in EFL Writing: A case of ChatGPT. *International Journal of Progressive Sciences and Technologies*, *39*(2), 357. https://doi.org/10.52155/ijpsat.v39.2.5516

Hidayatullah, E. (2024). Exploring Interactivity and Engagement: Improving Writing Skills with Chatgpt for Fun Learning. *Journal of Language Literature and Teaching*, *5*(3), 16–26. https://doi.org/10.35529/jllte.v5i3.16-26

Hong, Q.N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M., 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

Huawei, S., Aryadoust, V. (2022). A Systematic Review of Automated Writing Evaluation Systems. *Education and Information Technologies*, 28(1), 771–795. https://doi.org/10.1007/s10639-022-11200-7

Hwang, H., Chang, X., Sun, J. (2025). Generative AI is useful for Second Language Writing, but when, why, and for how long do Learners use it? *Journal of Second Language Writing*, *69*, 101230. https://doi.org/10.1016/j.jslw.2025.101230

Ingley, S.J., Pack, A. (2023). Leveraging AI Tools to develop the Writer rather than the Writing. *Trends in Ecology & Evolution*, *38*(9), 785–787. https://doi.org/10.1016/j.tree.2023.05.007

Jiang, Z., Xu, Z., Pan, Z., He, J., Xie, K. (2023). Exploring the role of Artificial intelligence in facilitating Assessment of Writing Performance in Second Language Learning. *Languages*, *8*(4), 247. https://doi.org/10.3390/languages8040247

Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, Jill Burstein. (2023). Rating Short L2 Essays on the CEFR Scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, Toronto, Canada, 576–584. https://aclanthology.org/2023.bea-1.49

Kim, J., Yu, S., Detrick, R., Li, N. (2025). Exploring Students' Perspectives on Generative AI-Assisted Academic Writing. Education and Information Technologies, 30(1), 1265-1300. https://doi.org/10.1007/s10639-024-12878-7

Kofinas, A.K., Tsay, C.H., Pike, D. (2025). The Impact of Generative AI on Academic Integrity of Authentic Assessments within a Higher Education Context. *British Journal of Educational Technology*, *56*(6), 2522–2549. https://doi.org/10.1111/bjet.13585

Kohnke, L. (2024). Exploring EAP Students' Perceptions of GenAI and Traditional Grammar-Checking Tools for Language Learning. *Computers and Education Artificial Intelligence*, *7*, 100279. https://doi.org/10.1016/j.caeai.2024.100279

Kohnke, L., Moorhouse, B.L., Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, 54(2), 537–550. https://doi.org/10.1177/00336882231162868

Koltovskaia, S., Rahmati, P., Saeli, H. (2024). Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement. *Journal of Second Language Writing*, *65*, 101130. https://doi.org/10.1016/j.jslw.2024.101130

Krakowski, S. (2025). Human-AI Agency in the Age of Generative AI. *Information and Organization*, *35*(1), 100560. https://doi.org/10.1016/j.infoandorg.2025.100560

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*, 100959. https://doi.org/10.1016/j.asw.2025.100959

Lee, S., Choe, H., Zou, D., Jeon, J. (2025). Generative AI (GenAI) in the language classroom: A systematic review. *Interactive Learning Environments*, *34*(1), 335–359. https://doi.org/10.1080/10494820.2025.2498537

Lee, Y.J., Davis, R.O., Lee, S.O. (2024). University Students' Perceptions of Artificial Intelligence-based tools for English Writing Courses. Online Journal of Communication and Media Technologies, 14(1), e202412. https://doi.org/10.30935/ojcmt/14195

Li, A.W., Collins, P. (2026). Formative Feedback across Sources: Student Perceptions and Writing Outcomes with instructor, peer, and AI-generated feedback. Reading and Writing, 1-23. https://doi.org/10.1007/s11145-026-10761-0

Li, C., Cui, H., Hagedorn, L. S. (2026). The Cognitive Impact of ChatGPT in Higher Education: A Systematic Review of Critical and Creative Thinking Outcomes. *Computers and Education Artificial Intelligence*, *10*, 100571. https://doi.org/10.1016/j.caeai.2026.100571

Li, S. (2025). Generative AI and Second Language Writing. *Digital Studies in Language and Literature*, 2(1), 122–152. https://doi.org/10.1515/dsll-2025-0007

Lin, S., Crosthwaite, P. (2024). The Grass is not Always Greener: Teacher vs. GPT-Assisted Written Corrective Feedback. *System*, *127*, 103529. https://doi.org/10.1016/j.system.2024.103529

Liu, C., Hou, J., Tu, Y., Wang, Y., Hwang, G. (2021). Incorporating a Reflective Thinking Promoting Mechanism into Artificial Intelligence-Supported English writing environments. *Interactive Learning Environments*, *31*(9), 5614–5632. https://doi.org/10.1080/10494820.2021.2012812

Liu, Z., Hwang, G., Chen, C., Chen, X., Ye, X. (2024). Integrating large language models into EFL writing instruction: effects on performance self-regulated learning strategies, and motivation. *Computer Assisted Language Learning*, 39(3), 466–490. https://doi.org/10.1080/09588221.2024.2389923

Lo, J., Wong, C., Ng, A., Wong, P., Cheung, D., Lai, P. (2025). Stretching AI's reach: Assessing an AI-Driven Feedback system for extended academic writing. Computers and Education: Artificial Intelligence, 100511. https://doi.org/10.1016/j.caeai.2025.100511

Lu, Q., Yao, Y., Xiao, L., Yuan, M., Wang, J., Zhu, X. (2024). Can ChatGPT effectively complement teacher assessment of undergraduate students' academic writing? *Assessment & Evaluation in Higher Education*, 49(5), 616–633. https://doi.org/10.1080/02602938.2024.2301722

Luo, Y., Xia, Y., Lu, X. (2025). A Systematic review of the role of Artificial intelligence in second language writing education. *Digital Studies in Language and Literature*, 2(2), 302–325. https://doi.org/10.1515/dsll-2025-0001

Lyu, W., Zhang, S., Chung, T., Sun, Y., Zhang, Y. (2025). Understanding the Practices, Perceptions, and (dis) Trust of Generative AI among Instructors: A Mixed-Methods Study in the US higher education. Computers and Education: Artificial Intelligence, 8, 100383. https://doi.org/10.1016/j.caeai.2025.100383

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

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

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), 2236469. https://doi.org/10.1080/2331186X.2023.2236469

Mhilli, O. (2023). Authorial Voice in Writing: A Literature Review. Social Sciences & Humanities Open, 8(1), 100550. https://doi.org/10.1016/j.ssaho.2023.100550

Minh, A.N. (2024). Leveraging ChatGPT for enhancing English writing skills and critical thinking in university freshmen. *Journal of Knowledge Learning and Science Technology ISSN 2959-6386 (Online)*, *3*(2), 51–62. https://doi.org/10.60087/jklst.vol3.n2.p62

Mizumoto, A., Eguchi, M. (2023). Exploring the Potential of using an AI Language Model for Automated Essay Scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mohamed, A. M. (2023). Exploring the potential of an AI-based Chatbot (ChatGPT) in enhancing English as a Foreign Language (EFL) Teaching: perceptions of EFL Faculty Members. *Education and Information Technologies*, *29*(3), 3195–3217. https://doi.org/10.1007/s10639-023-11917-z

Moorhouse, B.L., Wan, Y., Wu, C., Wu, M., Ho, T.Y. (2025). Generative AI Tools and Empowerment in L2 Academic Writing. *System*, *133*, 103779. https://doi.org/10.1016/j.system.2025.103779

Nguyen, L., Barrot, J. S. (2024). Detecting and assessing AI-generated and human-produced texts: The case of second language writing teachers. *Assessing Writing*, 62, 100899. https://doi.org/10.1016/j.asw.2024.100899

Ouzzani, M., Hammady, H., Fedorowicz, Z., Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1), 210. https://doi.org/10.1186/s13643-016-0384-4

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, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Page, M. J., McKenzie, J. E., Green, S. E., Forbes, A.B. (2013). An empirical investigation of the potential impact of selective inclusion of results in systematic reviews of interventions: study protocol. *Systematic Reviews*, *2*(1), 21. https://doi.org/10.1186/2046-4053-2-21

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., Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, Moher, D. (2021a). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), 89. https://doi.org/10.1186/s13643-021-01626-4

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., Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, Moher, D. (2021b). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Pfau, A., Polio, C., Xu, Y. (2023). Exploring the Potential of ChatGPT in Assessing L2 Writing Accuracy for Research Purposes. *Research Methods in Applied Linguistics*, 2(3), 100083. https://doi.org/10.1016/j.rmal.2023.100083

Poole, F.J., Coss, M.D. (2024). Can ChatGPT Reliably and Accurately Apply a Rubric to L2 Writing Assessments? The Devil is in the Prompt (s). *Journal of Technology & Chinese Language Teaching*, *15*(1).https://doi.org/10.35542/osf.io/3r2zb

Prakash, A., Aggarwal, S., Varghese, J.J., Varghese, J.J. (2025). Writing without Borders: AI and cross-Cultural Convergence in Academic Writing Quality. Humanities and Social Sciences Communications, 12(1), 1-11. https://doi.org/10.1057/s41599-025-05484-6

Ramadani, F., Revas, A., Tri Nugroho, R., Kartini, K., (2025). Artificial Intelligence (AI) Tools in Supporting Students Academic Writing Tasks: Benefits and Limitations. Jurnal Pendidikan dan Sastra Inggris, 5(3), https://doi.org/10.55606/jupensi.v5i3.6100

Ren, B., Dong, Y., Zhang, B. (2025). From Errors to Excellence: AI-Driven Scaffolding for Advancing L2 Writing Skills. Cogent Education, 12(1), 2588512. https://doi.org/10.1080/2331186X.2025.2588512

Richtig, G., Berger, M., Koeller, M., Richtig, M., Richtig, E., Scheffel, J., Maurer, M., Siebenhaar, F. (2023). Predatory journals: Perception, impact and use of Beall's list by the scientific community–A bibliometric big data study. *PLoS ONE*, 18(7), e0287547. https://doi.org/10.1371/journal.pone.0287547

Sahan, K., Kamaşak, R., Rose, H. (2023). The interplay of motivated behaviour, self-concept, self-efficacy, and language use on ease of academic study in English medium education. *System*, 114, 103016. https://doi.org/10.1016/j.system.2023.103016

Sar, E.B.O.H.K., Long, H.S. (2024). Exploring the use of ChatGPT as a Tool for Written Corrective Feedback in an EFL classroom. *The Journal of AsiaTEFL*, *21*(2), 397–412. https://doi.org/10.18823/asiatefl.2024.21.2.8.397

Saricaoglu, A., Bilki, Z. (2025). The capacity of ChatGPT-4 for L2 Writing Assessment: A closer look at Accuracy, Specificity, and Relevance. *Annual Review of Applied Linguistics*, *45*, 253–273. https://doi.org/10.1017/s0267190525100160

Seufert, S., Hartmann, P., Spirgi, L. (2025). Fostering Intelligent-TPACK through AI-Assistance: A Multi-Method Study in Pre-Service Teacher Education. Computers and Education Open, 100314. https://doi.org/10.1016/j.caeo.2025.100314

Shi, H., Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, *36*(2), 187–209. https://doi.org/10.1017/s0958344023000265

Shi, H., Chai, C.S., Zhou, S., Aubrey, S. (2025). Comparing the Effects of ChatGPT and Automated Writing Evaluation on Students' Writing and Ideal L2 Writing Self. *Computer Assisted Language Learning*, 1–28. https://doi.org/10.1080/09588221.2025.2454541

Song, C., Song, Y. (2023). Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, 14, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Steen-Utheim, A., Wittek, A.L. (2017). Dialogic Feedback and Potentialities for Student Learning. Learning, Culture and Social Interaction, 15, 18-30. https://doi.org/10.1016/j.lcsi.2017.06.002

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M.,Olson, C.B. (2024). Comparing the Quality of Human and ChatGPT Feedback of Students' Writing. *Learning and Instruction*, 91, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Strobl, C., Menke-Bazhutkina, I., Abel, N., Michel, M. (2024). Adopting ChatGPT as a Writing Buddy in the Advanced L2 Writing Class. *Technology in Language Teaching & Learning*, *6*(1), 1168. https://doi.org/10.29140/tltl.v6n1.1168

Su, Y., Lin, Y., Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, 57, 100752. https://doi.org/10.1016/j.asw.2023.100752

Tarchi, C., Zappoli, A., Ledesma, L.C., Brante, E.W. (2024). The use of ChatGPT in Source-Based Writing Tasks. *International Journal of Artificial Intelligence in Education*, *35*(2), 858–878. https://doi.org/10.1007/s40593-024-00413-1

Teng, M.F., Huang, J. (2026). Assessing ChatGPT feedback for EFL learners' engagement and writing performance. *International Journal of Applied Linguistics*. https://doi.org/10.1111/ijal.70155

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*, 103528. https://doi.org/10.1016/j.system.2024.103528

Tsao, J. (2025). Trajectories of AI policy in Higher Education: Interpretations, Discourses, and Enactments of Students and Teachers. Computers and Education: Artificial Intelligence, 100496. https://doi.org/10.1016/j.caeai.2025.100496

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

Van Niekerk, J., Delport, P.M., Sutherland, I. (2025). Addressing the use of Generative AI In Academic Writing. Computers and Education: Artificial Intelligence, 8, 100342. https://doi.org/10.1016/j.caeai.2024.100342

Wang, C. (2024). Exploring Students' Generative AI-Assisted Writing Processes: Perceptions and Experiences from Native and Nonnative English Speakers. *Technology Knowledge and Learning*, 30(3), 1825–1846. https://doi.org/10.1007/s10758-024-09744-3

Wang, C., Wang, Z. (2025). Investigating L2 writers' critical AI literacy in AI-assisted writing: An APSE model. *Journal of Second Language Writing*, 67, 101187. https://doi.org/10.1016/j.jslw.2025.101187

Wang, Y., Huang, J., Du, L., Guo, Y., Liu, Y., Wang, R. (2025). Evaluating Large Language Models as raters in Large-Scale Writing Assessments: A Psychometric Framework for Reliability and Validity. Computers and Education: Artificial Intelligence, 100481. https://doi.org/10.1016/j.caeai.2025.100481

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., Tate, T. (2023). The Affordances and Contradictions of AI-generated Text for Writers of English as a Second or Foreign Language. *Journal of Second Language Writing*, 62, 101071. https://doi.org/10.1016/j.jslw.2023.101071

Westbye, A. K., Eriksen, H., Blikstad-Balas, M. When AI Only Asks: How Question-Driven Dialogue Shapes Prewriting in the Classroom. In Frontiers in Education Frontiers, 11, 1740044. https://doi.org/10.3389/feduc.2026.1740044

Wiboolyasarin, W., Wiboolyasarin, K., Suwanwihok, K., Jinowat, N., Muenjanchoey, R. (2024). Synergizing Collaborative Writing and AI feedback: An Investigation into enhancing L2 Writing Proficiency in wiki-based Environments. *Computers and Education Artificial Intelligence*, *6*, 100228. https://doi.org/10.1016/j.caeai.2024.100228

Wiboolyasarin, W., Wiboolyasarin, K., Tiranant, P., Jinowat, N., Boonyakitanont, P. (2025). AI-driven Chatbots in Second Language Education: A Systematic Review of their Efficacy and Pedagogical Implications. *Ampersand*, *14*, 100224. https://doi.org/10.1016/j.amper.2025.100224

Wilson, J., Huang, Y. (2024). Validity of Automated Essay Scores for Elementary-Age English language Learners: Evidence of Bias? *Assessing Writing*, 60, 100815. https://doi.org/10.1016/j.asw.2024.100815

Xiao, Y., Zhi, Y. (2023). An exploratory study of EFL learners' use of ChaTGPT for language learning tasks: Experience and Perceptions. *Languages*, 8(3), 212. https://doi.org/10.3390/languages8030212

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, 28(11), 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

Yang, C., Li, R., Yang, L. (2025). Revisiting Trends in GenAI-Assisted Second Language Writing: Retrospect and Prospect. *Journal of Educational Computing Research*, 63(7–8), 1819–1863. https://doi.org/10.1177/07356331251367309

Yang, L., Li, R. (2024). ChatGPT for L2 learning: Current status and implications. System, 124, 103351. https://doi.org/10.1016/j.system.2024.103351

Yang, N., Gao, N., Zhang, T. (2025). Artificial Intelligence in EFL writing Enhancement. *International Journal of Web-Based Learning and Teaching Technologies*, *20*(1), 1–40. https://doi.org/10.4018/ijwltt.390442

Yu, J., Dai, X., Qiu, B., Guo, W., Wang, R., Na, M. (2026). AI Expectation Violations and Learner Engagement in EFL contexts: a cognitive-affective recovery Model. Frontiers in Psychology, 17, 1707116. https://doi.org/10.3389/fpsyg.2026.1707116

Zhan, Y., Yan, Z. (2025). Students' Engagement with ChatGPT Feedback: Implications for Student Feedback Literacy in the context of Generative Artificial Intelligence. *Assessment & Evaluation in Higher Education*, 1–14. https://doi.org/10.1080/02602938.2025.2471821

Zhang, R., Zou, D., Cheng, G. (2023). Chatbot-based Learning of Logical Fallacies in EFL Writing: Perceived Effectiveness in Improving Target Knowledge and Learner Motivation. *Interactive Learning Environments*, *32*(9), 5552–5569. https://doi.org/10.1080/10494820.2023.2220374

Zhang, Z., Aubrey, S., Huang, X., Chiu, T.K.F. (2025). The Role of Generative AI and Hybrid Feedback in Improving L2 Writing Skills: a Comparative Study. *Innovation in Language Learning and Teaching*, 1–19. https://doi.org/10.1080/17501229.2025.2503890

Zhang, Z., Hyland, K. (2023). Student Engagement with Peer Feedback in L2 writing: Insights from Reflective Journaling and Revising practices. *Assessing Writing*, *58*, 100784. https://doi.org/10.1016/j.asw.2023.100784

Zou, M., Huang, L. (2023). The impact of ChatGPT on L2 writing and expected responses: Voice from doctoral students. *Education and Information Technologies*, 29(11), 13201–13219. https://doi.org/10.1007/s10639-023-12397-x

Zuo, X., Zhu, R. (2025). Generative Artificial intelligence for L2 writing Feedback: Potentials and pitfalls. *US-China Education Review A*, 15(12), 845. https://doi.org/10.17265/2161-623x/2025.12.004

Has this article been screened for Similarity?

Yes

Conflict of interest

The Author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Declaration of AI Use

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) solely to assist with language editing, including grammar, spelling, sentence structure, and readability. The tool was not used to generate scientific content,

research data, analysis, interpretation and conclusions. All output was critically reviewed and edited by the authors, who take full responsibility for the final content of the manuscript.

About The License