Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Artificial Intelligence in EFL and ESP Academic Writing: A Bibliometric and Critical Systematic Review of Pedagogical Gains and Assessment Risks

Zulfiya Khabirova Anvarbekovna

zulya.uz@gmail.com

AJOU University in Tashkent https://orcid.org/0009-0008-6726-6934

Research Article

Keywords: generative artificial intelligence, ChatGPT, automated writing evaluation, EFL academic writing, English for specific purposes, AI text detection, bibliometric analysis, VOSviewer, systematic review

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11627257/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Artificial Intelligence in EFL and ESP Academic Writing: A Bibliometric and Critical Systematic Review of Pedagogical Gains and Assessment Risks

Zulfiya Khabirova Anvarbekovna[1]*

[1] AJOU University in Tashkent, Tashkent, Uzbekistan; https://orcid.org/0009-0008-6726-6934

Email: zulya.uz@gmail.com

Abstract

Background: Artificial intelligence (AI) writing tools, from automated writing evaluation (AWE) and paraphrasers to generative chatbots, are now routine in the academic writing of learners of English as a foreign language (EFL) and English for specific purposes (ESP). The literature has expanded faster than it has been integrated: studies of learning gains, feedback, teacher attitudes and academic integrity have developed largely in parallel.

Aim: This review maps the structure and evolution of research on AI in EFL/ESP academic writing and critically evaluates what the evidence does and does not show, with particular attention to the relation between pedagogical benefit and assessment risk.

Method: A hybrid bibliometric–systematic design informed by PRISMA 2020 was applied to OpenAlex. A three-block title/abstract search (2019–2026) and a journal-scope supplementary search yielded 4,846 records; transparent rule-based screening retained 2,612 journal documents. VOSviewer 1.6.21 was used for term co-occurrence (130 terms), overlay, reference co-citation (200 references) and country co-authorship (36 countries) analyses. Bibliometric results were integrated with abstract-level coding of 2,448 records and close reading of the most influential and gap-relevant studies.

Results: Output grew from 6 documents in 2019 to 1,261 in 2026 (to 4 November); 79.5% appeared in 2025–2026. Six thematic clusters emerged: feedback efficacy, integrity and governance, assessment and measurement, agency and authorship, scholarly publishing by non-native writers, and technology acceptance. Only 16.1% of term-link strength connected the pedagogical clusters with the integrity/authorship clusters. The intellectual base comprised an AWE and corrective-feedback foundation, a 2023 ChatGPT opportunity–risk wave and an emerging GenAI-feedback and assessment front. Of 492 experimental or quasi-experimental studies, 9.8% reported delayed or longitudinal measures, 5.7% examined transfer to unaided writing and 3.7% were situated in ESP; only 61 records (2.5%) addressed AI-text detection.

Conclusions: The field documents robust, short-term text-level gains but has rarely tested durable learning, and it has not connected those gains with the assessment regimes that judge AI-assisted texts. We conceptualise this disconnection as a mediated-accuracy paradox: the linguistic polish that AI makes available to L2 writers is also a feature on which style-sensitive detectors may act. An integrative framework, five propositions and a research agenda are proposed.

Keywords: generative artificial intelligence; ChatGPT; automated writing evaluation; EFL academic writing; English for specific purposes; AI text detection; bibliometric analysis; VOSviewer; systematic review

1. Introduction

Academic writing has long been the most demanding skill for learners of English as a foreign language (EFL). Recent syntheses and descriptive reviews continue to report persistent difficulties with grammar, punctuation, lexical choice, coherence and the organisation of ideas, compounded by limited explicit instruction and first-language transfer (Raheem and Ghafar, 2025). In English for specific purposes (ESP), the challenge is sharper because writers must also master discipline-specific genres whose conventions are rarely made explicit (Hyland, 2025; Swales, 1990). Into this space have entered successive generations of AI writing technologies: AWE systems and grammar checkers, AI paraphrasers such as QuillBot and, since the public release of ChatGPT in November 2022, large language model (LLM) chatbots that can draft, revise and comment on text on demand (Banh and Strobel, 2023; Feuerriegel et al., 2023; Kasneci et al., 2023; Ali et al., 2024).

Empirical studies report that these tools can improve the accuracy, organisation and lexical range of L2 texts and can raise writers' confidence (Song and Song, 2023; Alkamel and Alwagieh, 2024; Marzuki et al., 2023; Mahapatra, 2024). A recent quasi-experiment with Indonesian undergraduates, for example, found significantly higher post-test academic writing scores in a ChatGPT-assisted group than in a conventionally taught group (Amanda et al., 2026), and a mixed-methods study of business English students reported near-unanimous perceptions that QuillBot improved grammatical accuracy, efficiency and confidence (Rozi et al., 2026). At the same time, ESP teachers' engagement with generative AI (GenAI) ranges from technophilia to technophobia and is shaped by colleagues, student expectations and unclear institutional policies (Alrashed et al., 2026). Concerns about over-reliance, authorship and academic integrity are widespread (Barrot, 2023; Cotton et al., 2023; Dergaa et al., 2023), and institutions increasingly rely on AI-text detectors whose outputs are sensitive to linguistic style. A large paired-document study of professionally edited manuscripts by non-native writers showed that detector false-positive rates on human-written text ranged from 0% to 100% across 13 detectors and that editing alone shifted detector scores (Park et al., 2026), extending earlier evidence that GPT detectors disproportionately flag non-native English writing (Liang et al., 2023).

Read together, these five recent studies, which form the seed set of the present review, expose a tension that the literature has rarely addressed directly. The same AI-mediated polish that studies of learning outcomes treat as a benefit can become a liability when texts are judged by integrity regimes that infer authorship from style. Whether the field has recognised this

tension depends on how its research streams are connected, and that is an empirical question about the structure of the literature as well as a conceptual one.

Existing reviews provide important but partial answers. Systematic reviews have synthesised AI-based automated written feedback (Shi and Aryadoust, 2024), ChatGPT as a writing assistant in higher education (Imran and Almusharraf, 2023) and AI applications in higher education more broadly, the last concluding that the field lacked critical reflection on risks and weak links to pedagogical theory (Zawacki-Richter et al., 2019). A scientometric analysis of ChatGPT in academic writing covered only 2022–2023 and was not specific to L2 writers (Lendvai, 2025), and descriptive reviews of EFL writing have been small in scale (Raheem and Ghafar, 2025). To our knowledge, no study has combined a large-scale science map of AI-mediated L2 academic writing with a critical appraisal of the evidence, extended the period to the post-2023 surge, or examined whether the pedagogical and integrity strands of the field speak to each other.

This review addresses that gap. It combines a VOSviewer-based bibliometric mapping (van Eck and Waltman, 2009) of 2,612 OpenAlex documents with a critical literature analysis to answer four research questions (RQs):

- **RQ1.** How has research on AI in EFL/ESP academic writing developed in terms of output, publication venues and international collaboration?
- **RQ2.** What thematic structure, intellectual base and temporal evolution characterise the field?
- **RQ3.** What does a critical synthesis of the evidence show about the effects of AI on L2 academic writing, the quality of AI feedback, teacher and learner engagement, and integrity and assessment?
- **RQ4**. What gaps follow from integrating the bibliometric and critical analyses, and what framework and research agenda can address them?

The review makes three contributions. Empirically, it provides a reproducible map of a fast-moving field, with all search strings, screening decisions and VOSviewer inputs released. Analytically, it quantifies the weak integration between research on pedagogical gains and research on integrity and authorship, and it profiles the design features of intervention studies. Conceptually, it proposes the mediated-accuracy paradox and an integrative framework with five testable propositions that connect L2 writers, AI mediation, learning outcomes and assessment validity.

2. Theoretical framework

Four complementary perspectives frame the review. Each explains a different part of what happens when an L2 writer composes with AI, and together they define the constructs that were coded and mapped (Section 3) and the propositions derived in Section 6.

2.1 Sociocultural mediation and scaffolding

Sociocultural learning theory can be outlined as follows: Higher mental functions emerge in activity supported by tools and signs, learning takes place when support is provided in ZPD (Vygotsky, 1978). In its original definition, scaffolding is support that is contingent on the learner's current level of performance and is faded away as competence increases (Wood et al., 1976). This is a key aspect of AI writing tools. An LLM can be used as a scaffold, to elicit ideas, model a move or explain an error, or be a substitute that does the writing for the student. This idea is implicit in studies that see ChatGPT as a mediator of academic writing (Fathi and Rahimi, 2024) and as a scaffold for argumentation (Guo et al., 2022; Su et al., 2023); however, the theory suggests a test that is infrequently given: mediated performance becomes learning when it is internalized and transferable to unaided performance. The other side of the coin is the risk associated with cognitive offloading, which is the external resource used to reduce the internal cognitive load (Risko and Gilbert, 2016).

2.2 Feedback and learner engagement

The second lens is provided by feedback research. Feedback brings two key aspects to a learner's task: first, where the learner is going and second, how the learner is progressing and what he/she needs to do next. (Hattie and Timperley, 2007). As for L2 writing, the impact of feedback can be understood through writers' behavioural, cognitive and affective responses to the feedback (Ellis, 2010); and writers' abilities in interpreting and responding to the feedback (Carless and Boud, 2018). The results of the research on AWE revealed that automated feedback can facilitate revision and accuracy, and that the effectiveness depends on the accuracy of the feedback, the learner's engagement and the teacher's integration of the tool (Li et al., 2015; Ranalli, 2018; Koltovskaia, 2020; Hyland and Hyland, 2006). This is a baseline with which to assess claims regarding the feedback of GenAI.

2.3 Genre and academic literacies

Academic and professional writing are not neutral skills, but social practices that are part of a disciplinary community. Both genre theory (Swales, 1990) and ESP writing pedagogy (Hyland, 2025) highlight the communicative functions and rhetorical strategies that are assumed to be used by academic or workplace discourse communities. The academic literacies tradition adds that student writing is bound up with identity, power and institutional expectations (Lea and Street, 1998). These types of perspectives emphasize things that are at the surface level that is not addressed in the accuracy measures: voice, stance, ownership, appropriation of disciplinary conventions. They also explain why ESP contexts, where genre demands are explicit and high-stakes, are a critical test case for AI-mediated writing (Rozi et al., 2026; Alrashed et al., 2026). The cognitive model of writing as recursive planning, translating and reviewing (Flower and Hayes, 1981) complements this view by locating where AI assistance enters the composing process.

2.4 Technology acceptance and teacher cognition

A substantial body of work explains AI use through the technology acceptance model and the unified theory of acceptance and use of technology, in which perceived usefulness, ease of use, social influence and facilitating conditions predict behavioural intention (Davis, 1989; Venkatesh et al., 2003); digital literacy has also been shown to shape these perceptions and, through them, behavioural intention (Akram et al., 2023). In addition, anxiety and digital literacy and institutional support influence teacher acceptance or rejection of GenAI, as shown in teacher-focused studies (Verano-Tacoronte et al., 2025; Taheri et al., 2025; Qadir and Omar, 2025). In the case of ESP, teachers' attitudes can shift from technophile to technophobe, and are negotiated with colleagues and students (Alrashed et al., 2026). Acceptance models are models of intention, not models of learning or fairness, though, and the fact that they do so much have been the focus of the literature is itself a finding that should be investigated.

2.5 Assessment validity and fairness

The fourth lens is about inferences based on AI-generated texts. The validity of inferences made by institutions based on text features relies on whether or not features are specific to machine generation. There is evidence to the contrary. The results indicate that most TOEFL essays written by non-native speakers were misclassified as AI-generated by GPT detectors

(Liang et al., 2023); an analysis of 14 detectors found that none were accurate or reliable (Weber-Wulff et al., 2023); and the output of detectors is affected by the length of the text, genre, and hybrid human–AI authorship (Fraser et al., 2025; Hadra et al., 2026). Because proficiency in L2 writing is itself associated with measurable cohesion and lexical sophistication (Crossley and McNamara, 2010), detectors that respond to such features risk confounding language development, editing and AI use (Park et al., 2026). False positives therefore carry consequences for individuals that are not only technical but ethical (Giray, 2024).

2.6 Integrating the lenses: the mediated-accuracy problem

The four lenses converge on one problem. Sociocultural and feedback theories predict that AI support improves learning only when the writer engages with it actively and the support fades; genre and academic-literacies theories predict that what matters is the writer's appropriation of disciplinary voice; and assessment theory warns that the textual evidence used to judge authorship is itself shaped by the support received. We use the term mediated accuracy for the improved surface quality of a text produced with AI assistance, and we ask whether the literature distinguishes mediated accuracy from learning and whether it recognises the assessment consequences of mediated accuracy. These questions structure the coding dictionaries, the interpretation of the science maps and the framework presented in Section 6.

3. Methodology

3.1 Research design

The study used a hybrid bibliometric–systematic review design. Bibliometric mapping was used to describe the structure and evolution of the field at scale (Donthu et al., 2021; Zupic and Čater, 2014), and a critical evidence synthesis was used to evaluate the substance and quality of the findings. Reporting follows the PRISMA 2020 statement where applicable to a bibliometric corpus (Page et al., 2021). Figure 1 summarises the identification, screening and analysis stages. Python scripts implemented each step, and all intermediate files are provided in Supplementary File S1 so that every number reported below can be reproduced.

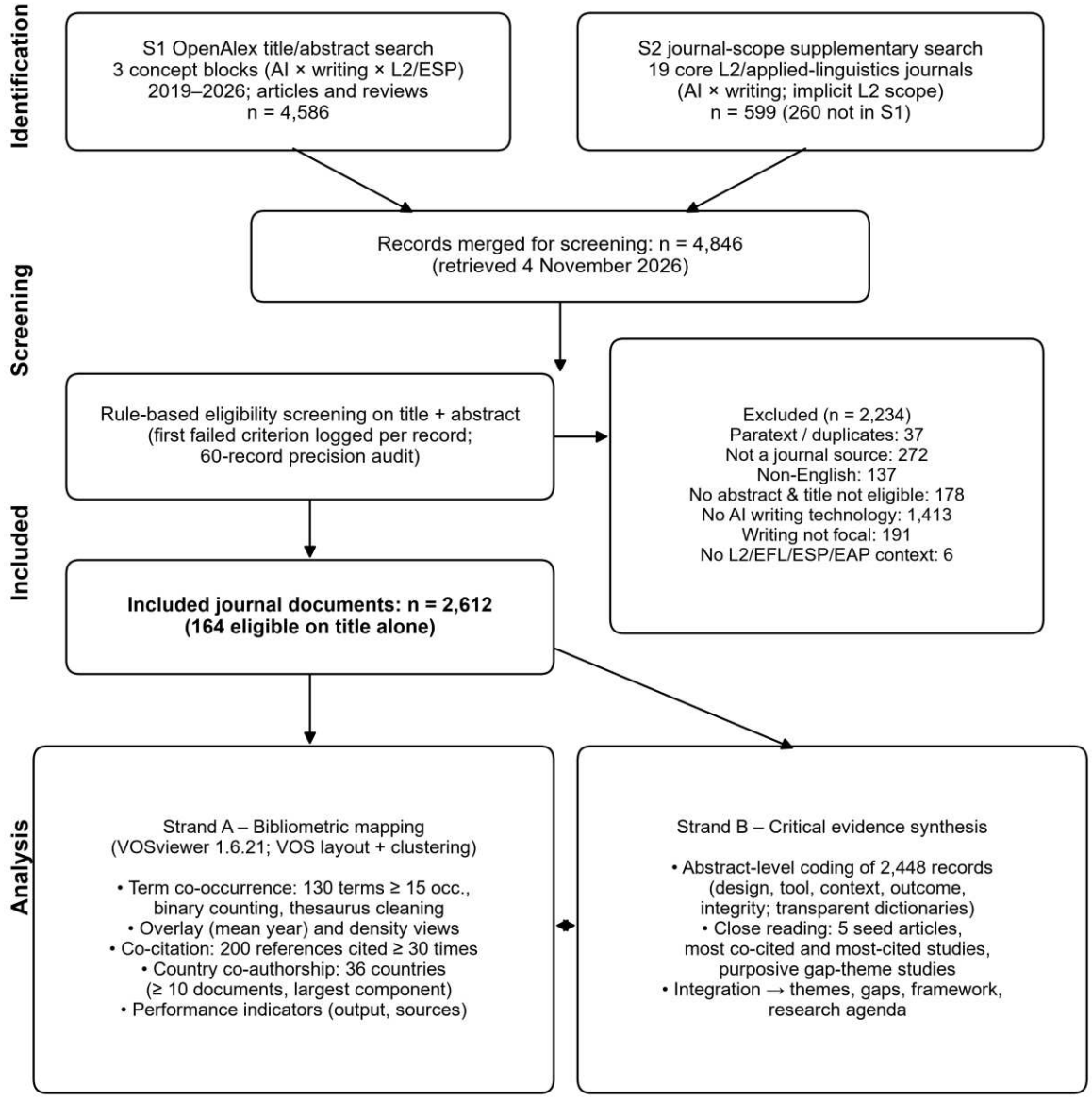


Figure 1. Identification, screening and analysis flow of the hybrid bibliometric–systematic review (adapted from PRISMA 2020). Exclusion reasons record the first failed criterion for each record.

3.2 Database selection

OpenAlex was selected as the data source for three reasons. First, it is an open bibliographic index that covers journals from regions that are under-represented in Web of Science and Scopus but are, as Section 4 shows, central to this field (notably Indonesia, Malaysia, Saudi Arabia, Pakistan and Uzbekistan) (Priem et al., 2022; Alperin et al., 2024). Second, its reference lists permit co-citation analysis at a coverage comparable to the proprietary databases (Culbert et al., 2025). Third, its records and query syntax are openly available (CC0), so the corpus can be audited and re-run, which proprietary exports do not allow. Two

known limitations were addressed in the design: abstracts are missing for some publishers, notably Elsevier, and metadata on language and document type are occasionally wrong (Sections 3.3–3.5).

3.3 Search strategy

The main search (S1) combined three concept blocks with Boolean AND in the title and abstract field: AI writing technologies, academic and L2 writing, and second-language or specific-purposes populations (Table 1). Terms were derived from the seed articles, from the keywords of highly cited studies and from iterative piloting. Because many core L2 writing journals publish studies whose titles do not name the L2 population, and because OpenAlex lacks abstracts for several of them, a supplementary journal-scope search (S2) applied the AI and writing blocks to 19 core applied-linguistics and L2 journals (e.g., Assessing Writing, Journal of Second Language Writing, System, Computer Assisted Language Learning, ReCALL, Language Learning & Technology, Journal of English for Academic Purposes and English for Specific Purposes). Searches were limited to articles and reviews published from 2019, the year in which transformer-based text generation entered public discussion, to the retrieval date (4 November 2026).

Table 1. Search strategy (OpenAlex, title_and_abstract.search; retrieved 4 November 2026).

Block	Search terms (combined with OR)	Records
S1-A AI writing technologies	"artificial intelligence", "generative AI", "generative artificial intelligence", GenAI, ChatGPT, "large language model(s)", chatbot(s), "AI-powered", "AI-assisted", "AI-driven", "AI-based", "AI tools", "AI writing", "AI-generated", "automated writing evaluation", Grammarly, QuillBot	
S1-B Writing	"academic writing", "L2 writing", "second language writing", "EFL writing", "ESL writing", "writing skill(s)", "writing instruction", "writing performance", "essay writing", "argumentative writing", "written corrective feedback", "writing feedback", "writing proficiency", "writing quality", "writing class(es)", "writing course"	
S1-C L2/ESP population	EFL, ESL, "English as a foreign/second language", "second language", L2, "non-native", nonnative, ESP, "English for specific purposes", "English for academic purposes", EAP, "(English) language learners", "foreign language"	
S1 = A AND B AND C	type: article | review; publication year 2019–2026	4,586
S2 journal scope	Block A AND (writing, writers, essay(s), "written corrective	599 (260 new)

	feedback", "writing feedback") in 19 core L2/applied-linguistics journals	
Total	Unique records merged for screening	4,846

Note. Full Boolean strings and journal identifiers are given in Supplementary File S1 (oa_common.py).

3.4 Eligibility criteria

Eligibility criteria were defined before screening and operationalised as ordered rules (Table 2). Records were included if they were English-language journal articles or reviews with a usable abstract in which (a) an AI writing technology was a focal object, (b) writing was a focal skill and (c) the population or context involved L2, EFL, ESL, ESP or EAP writers. Focality was required to exclude records that mention AI or writing only in passing: AI had to appear in the title or at least twice in the abstract, and writing had to appear in the title or be a recurrent focus of the abstract. Records without an abstract were retained only if the title alone satisfied all three relevance criteria, which prevented the systematic loss of studies from publishers whose abstracts OpenAlex does not hold.

Table 2. Inclusion and exclusion criteria applied in order (first failed criterion logged).

Criterion	Inclusion	Exclusion (records excluded)
Document	Journal article or review	Paratext, errata, retractions, duplicates by DOI or normalised title (37); repository, preprint, conference or missing source (272)
Language	English (metadata and function-word and script check)	Non-English (137)
Information	Abstract ≥ 50 words, or title alone eligible	No usable abstract and title not eligible (178)
AI focus	AI writing technology in title or ≥ 2 abstract mentions	No AI writing technology or not focal (1,413)
Writing focus	Writing named in title or recurrent in abstract	Writing not focal (191)
Population	L2/EFL/ESL/ESP/EAP writers or core L2 journal scope	No L2 context (6)
Result	2,612 included (164 on title alone)	2,234 excluded

Note. Regular expressions for each rule are documented in screen.py (Supplementary File S1).

3.5 Screening, data cleaning and quality assurance

Rule-based screening was chosen over manual title screening because the volume of records (4,846) and the speed of publication made a transparent, repeatable procedure preferable to an undocumented one. Each record was tested against the criteria in a fixed order and the first failed criterion was logged, so that the exclusion reason is available for every record. Duplicates were removed by DOI and by normalised title. Because OpenAlex language tags are occasionally wrong, records were also excluded when their abstracts lacked English

function words or their titles were not in Latin script. To estimate precision, a random sample of 60 included records (seed 2026) was drawn and judged by the first author against the eligibility criteria, using titles and, for eight unclear cases, abstracts. Of the 59 records that remained in the final corpus, 56 (94.9%) were fully relevant and three (5.1%) partly relevant (e.g., studies of broader L2 proficiency in which writing was one of several skills); none was irrelevant. Recall was checked against a benchmark of known core studies, including the seed articles indexed in OpenAlex and highly cited studies such as Barrot (2023), Su et al. (2023), Song and Song (2023) and Mizumoto and Eguchi (2023); all were retrieved, the second and fourth only through S2, which confirms the value of the supplementary search.

For the term analysis, a thesaurus file merged spelling variants and synonyms (e.g., GenAI, generative AI and generative artificial intelligence; AWE and automated writing evaluation) and removed generic, statistical and publisher boilerplate terms (e.g., study, copyright, independent sample t test). The thesaurus was refined over three rounds by inspecting the resulting maps, a standard practice in co-word analysis. For co-citation, cited references were matched by OpenAlex identifier, labelled by first-author surname and year, and three book records lacking author metadata were relabelled manually (logged in reference_cleaning_log.tsv).

3.6 Bibliometric techniques

Performance analysis described annual output, sources, countries and citations. Science mapping used VOSviewer 1.6.21, which applies the VOS mapping technique and a unified modularity-based clustering algorithm to normalised co-occurrence matrices (van Eck and Waltman, 2009; Waltman et al., 2010). Four networks were analysed (Table 3). Term co-occurrence analysis identified research themes from noun phrases in titles and abstracts; an overlay visualisation coloured each term by the mean publication year of the documents in which it occurred, revealing emerging topics; and a density visualisation showed concentrations of research attention. Reference co-citation analysis (Small, 1973) identified the intellectual base, that is, groups of works that the corpus cites together. Country co-authorship analysis described collaboration based on institutional affiliations. VOSviewer was run from the command line with fixed parameters so that layouts and clusters are reproducible.

Table 3. VOSviewer analyses and parameters.

Analysis	Unit / counting	Threshold and items	Clustering
Term co-occurrence (Fig. 3)	Noun phrases in titles and	≥ 15 occurrences; 130 most	Resolution 1.2; min. cluster

	abstracts; binary counting; thesaurus applied	relevant terms	size 8; 6 clusters
Overlay (Fig. 4)	Mean publication year of documents per term	Same map; colour scale 2024.6–2025.8	—
Density (Fig. S1, S2)	Item density	Same maps as Figs 3 and 5	—
Reference co-citation (Fig. 5)	Cited references (OpenAlex IDs); full counting	Cited by ≥ 30 corpus documents; 200 references	Resolution 1.0; min. cluster size 8; 3 clusters
Country co-authorship (Fig. 6)	Countries of author affiliations; full counting	≥ 10 documents; largest connected component; 36 countries	Resolution 1.0; 6 clusters

Note. Common settings: attraction 2, repulsion 1, merge small clusters, curved lines. Scripts: run_vosviewer.py.

3.7 Critical evidence synthesis and abstract-level coding

Science maps describe structure but not the quality or direction of findings. The critical synthesis therefore had two components. First, all 2,448 included records with abstracts were coded with transparent regular-expression dictionaries for research design (experimental or quasi-experimental, survey/structural equation modelling, qualitative, review, textual or process data), technology (GenAI/LLM, AWE, Grammarly, QuillBot, machine translation), context (ESP, EAP and research writing, school), population (teachers), outcomes (affective, critical thinking and argumentation, agency and authorship), integrity and detection, and methodological features relevant to learning claims (delayed or longitudinal measurement; transfer to unaided writing). Dictionary coding is a coarse instrument: it indicates whether a feature is reported in an abstract, not whether it was studied well, and it may under-detect features reported only in full texts. Coding shares are therefore interpreted as indicators of research attention rather than exact prevalence. Second, a close reading was undertaken of the five seed articles, the most highly co-cited references in each co-citation cluster, the most cited corpus documents and purposively selected studies addressing the gap themes (detection, ESP, longitudinal designs and agency). Findings were synthesised thematically (Braun and Clarke, 2006) and organised by the bibliometric clusters.

3.8 Integration of bibliometric and critical analyses

Integration proceeded in three steps. Bibliometric clusters were first interpreted with the help of the studies that define them; the critical synthesis then assessed what each cluster's evidence supports; finally, cross-cluster link strengths and coding indicators were used to locate gaps, defined as combinations of constructs that theory identifies as important (Section 2) but that the network shows to be weakly connected or the coding shows to be rarely studied. The share of link strength between clusters was calculated from the VOSviewer network files, and the pedagogical block (feedback efficacy; assessment and measurement)

was compared with the integrity and authorship block (integrity and governance; agency and authorship; scholarly publishing).

4. Results

4.1 Growth, sources and collaboration (RQ1)

The 2,612 included documents comprised 2,521 articles and 91 reviews published in 1,171 sources by authors affiliated with 98 countries; together they had received 37,273 citations in OpenAlex (mean 14.3 per document). Output was negligible before ChatGPT (6 documents in 2019, 13 in 2020, 31 in 2021 and 45 in 2022) and then grew steeply to 110 in 2023, 331 in 2024, 815 in 2025 and 1,261 in 2026 to 4 November (Fig. 2a). More than three quarters of the corpus (79.5%) was published in 2025–2026, which means that most of the field is too recent to have accumulated citations or replications.

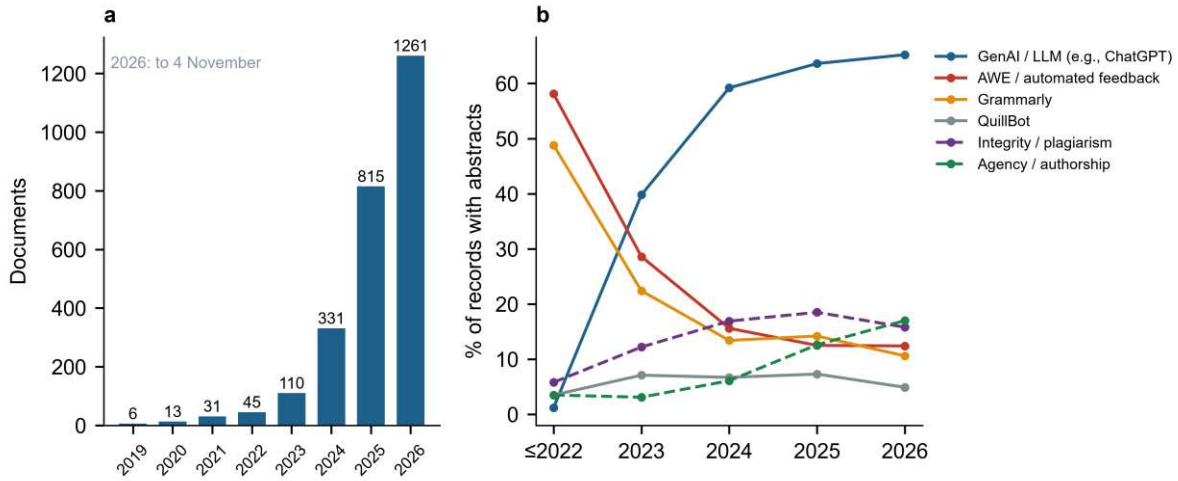


Figure 2. (a) Annual number of included documents, 2019–2026 (2026 to 4 November). (b) Share of records with abstracts mentioning selected technologies and themes by year (≤ 2022 pooled). Source: abstract-level coding of 2,448 records.

The technological object of research changed abruptly (Fig. 2b). Before 2023, AWE or automated feedback appeared in 58.1% of abstracts and Grammarly in 48.8%; by 2026 these shares had fallen to 12.4% and 10.6%, while GenAI or LLMs rose from 1.2% to 65.2%. Mentions of agency and authorship rose from 3.5% to 17.0%, and mentions of integrity or plagiarism from 5.8% to a peak of 18.5% in 2025. The field, in other words, did not accumulate on its AWE foundations so much as pivot to a new technology, a point taken up in Section 5.

Publication was widely dispersed. The most productive sources were Assessing Writing (51 documents; 3,216 citations), Computer Assisted Language Learning (47; 2,580), Innovation in Language Learning and Teaching (41; 698), System (35; 1,244) and Journal of Second Language Writing (27; 1,197) (Table 4). The high citation counts of these specialist journals contrast with several high-volume general or regional outlets whose documents were rarely cited, suggesting a two-tier structure in which a small set of journals shapes the intellectual base while a large periphery reproduces it.

Table 4. Most productive sources and countries (all included documents).

Source	Docs	Cites	Country	Docs	Cites
Assessing Writing	51	3,216	Indonesia	536	3,650
Computer Assisted Language Learning	47	2,580	China	416	9,819
Innovation in Language Learning and Teaching	41	698	Malaysia	156	1,088
Int. J. of Research and Innovation in Social Science	36	25	Saudi Arabia	154	1,984
System	35	1,244	United States	150	4,946
Arab World English Journal	32	303	Hong Kong	107	4,569
Journal of Second Language Writing	27	1,197	Türkiye	85	919
SAGE Open	24	120	Vietnam	75	740
IDEAS Journal on ELT and Learning, Linguistics and Literature	23	16	Pakistan	68	156
Frontiers in Psychology	23	1,510	Iran	56	2,113

Note. Countries are counted once per document for any author affiliation; citations are OpenAlex counts at retrieval.

Geographically, the field is led by Asia and the Middle East. Indonesia contributed the most documents (536), followed by China (416), Malaysia (156), Saudi Arabia (154) and the United States (150). Volume and visibility diverged sharply: Chinese documents averaged 23.6 citations and Hong Kong documents 42.7, compared with 6.8 for Indonesian documents. The co-authorship map (Fig. 6) shows six collaboration clusters. China had the highest total link strength (196), followed by Hong Kong (88) and the United States (73), whereas Indonesia, despite producing 20.5% of the corpus, had a total link strength of only 68, indicating predominantly domestic authorship. A Gulf–South Asian cluster (Malaysia, Saudi

Arabia, Pakistan, Iran, Uzbekistan, India, Oman, Egypt and others) and a Southeast Asian cluster centred on Indonesia, Vietnam and Thailand were linked to each other and to China, Hong Kong and the United Kingdom by comparatively thin ties.

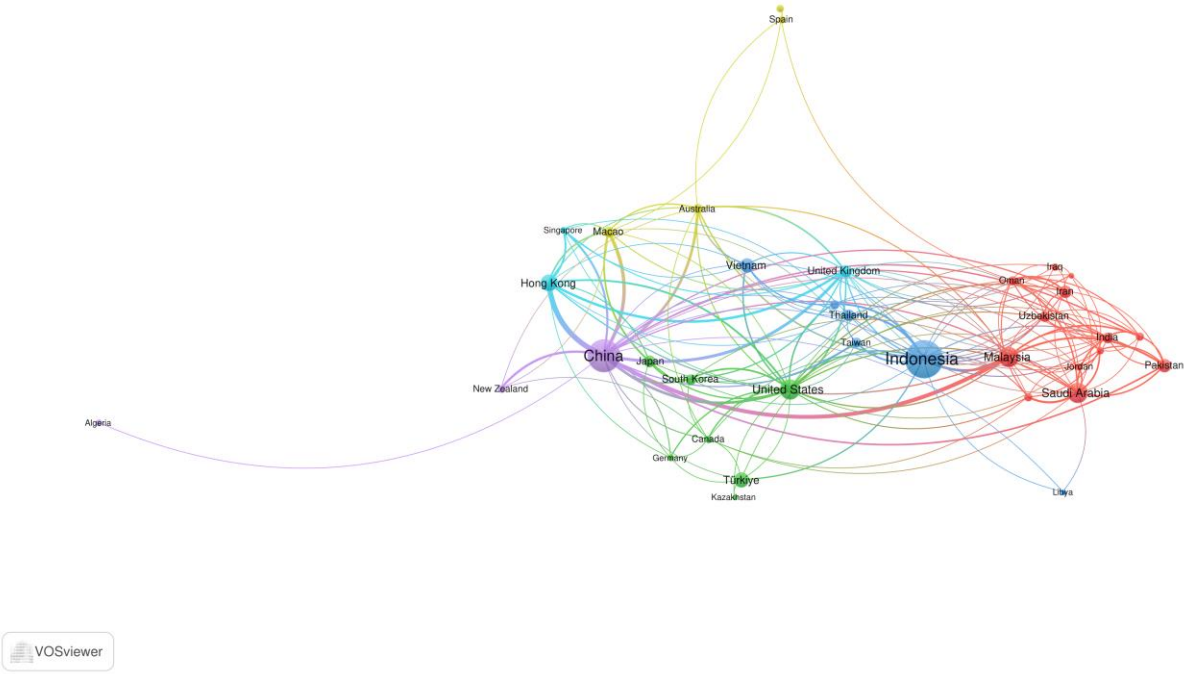


Figure 6. Country co-authorship network (VOSviewer; 36 countries with ≥ 10 documents, largest component). Node size: documents; line thickness: co-authored documents; colour: cluster.

4.2 Thematic structure (RQ2)

The term co-occurrence map (Fig. 3) contains 130 terms in six clusters (Table 5). The map is visibly bipolar. On the right, research on feedback efficacy (T1) and assessment and measurement (T3) forms a dense pedagogical region; on the left, integrity and governance (T2), agency and authorship (T4) and scholarly publishing by non-native writers (T5) form a second region; technology acceptance (T6) sits between them at the top.

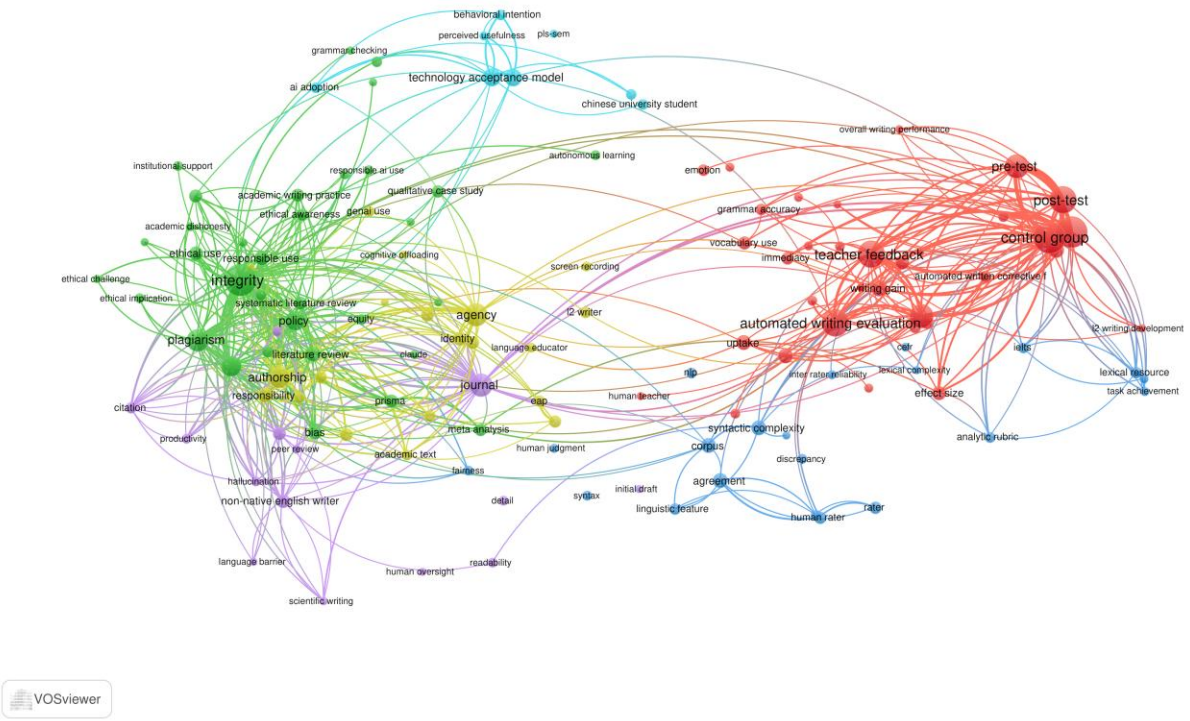


Figure 3. Term co-occurrence network (VOSviewer; 130 terms with ≥ 15 occurrences, binary counting). Node size: occurrences; colour: cluster (T1 red, T2 green, T3 blue, T4 yellow, T5 purple, T6 cyan).

Table 5. Thematic clusters of the term co-occurrence map.

Cluster (terms; mean year)	Most frequent terms	Interpretation
T1 Feedback efficacy and engagement (34; 2024.97)	control group, experimental group, post-test, automated writing evaluation, teacher feedback, corrective feedback, writing gain, uptake, effect size, affective/emotional engagement	Intervention studies comparing AI, AWE and teacher feedback on writing scores and engagement
T2 Integrity, ethics and governance (33; 2025.34)	integrity, plagiarism, guideline, policy, responsible use, bias, ethical use, institutional policy, overdependence, academic dishonesty, data privacy	Policy-oriented, often qualitative or review work on responsible use and misconduct
T3 Assessment and linguistic measurement (21; 2025.20)	corpus, agreement, syntactic complexity, human rater, automated essay scoring, IELTS, analytic rubric, NLP, fairness, inter-rater reliability, lexical sophistication	AI as scorer and the linguistic profile of AI-assisted or AI-generated text
T4 Agency, authorship and critical AI literacy (20; 2025.59)	agency, authorship, identity, verification, stance, critical AI literacy, ownership, EAP, L2 writer, authorial voice, screen recording, cognitive offloading	The newest front: who writes, and how L2 writers critically regulate AI use
T5 Scholarly publishing and non-native writers (14; 2025.11)	journal, non-native English writer, transparency, citation, hallucination, productivity, peer review, scientific writing, language barrier	AI in research publication by multilingual scholars
T6 Technology acceptance (8; 2025.40)	technology acceptance model, perceived ease of use, behavioural intention, AI adoption, structural equation modelling, perceived usefulness, PLS-SEM	Survey-based models of intention to use AI

Note. Mean year: occurrence-weighted mean of the terms' average publication year.

Two features of this structure are central to the argument of the review. First, the clusters are poorly integrated across the pedagogical and integrity regions (Fig. 7a). T1 directed 55.9% of its link strength to itself, the highest self-share of any cluster, and only 12.1% to integrity and governance, 7.6% to agency and authorship and 5.5% to scholarly publishing. Across the whole network, links within the pedagogical block accounted for 33.2% of link strength and links within the integrity–authorship block for 44.5%, but links between the two blocks for only 16.1%. Studies that measure what AI does for L2 texts and studies that discuss what AI means for authorship and integrity thus rarely share vocabulary. Second, the terms that would bridge the two regions sit in different clusters: fairness belongs to the assessment cluster (T3), whereas non-native English writer belongs to the publishing cluster (T5). The concern that detection may be unfair to L2 writers is therefore not connected, at the level of the field's vocabulary, with the pedagogical studies in which L2 writers' texts are improved by AI.

4.3 Temporal evolution

The sequence is clearly seen on the overlay map (Fig. 4). The mean years (T1, 2017.95) for terms relating to AWE, corrective feedback, uptake and engagement are the earliest, whereas the publishing and assessment clusters have the next most recent mean years (T2, 2024.79). The most recent terms focus on the agency and authorship cluster (T4, 2025.59), such as critical AI literacy, ownership, authorial voice, verification, cognitive offloading and screen recording. The change is from the question of whether AI "improves" text to the question of how writers control AI and whose writing it is. The use of process-oriented data (screen recordings, interaction logs) has grown in this front, from 10% of abstracts in 2023 to 10.8% in 2026, as well as the number of mentions of delayed or longitudinal designs, which increased from 1.0% to 10.8% in the abstracts. This indicates methodological maturation, albeit limited in absolute numbers.

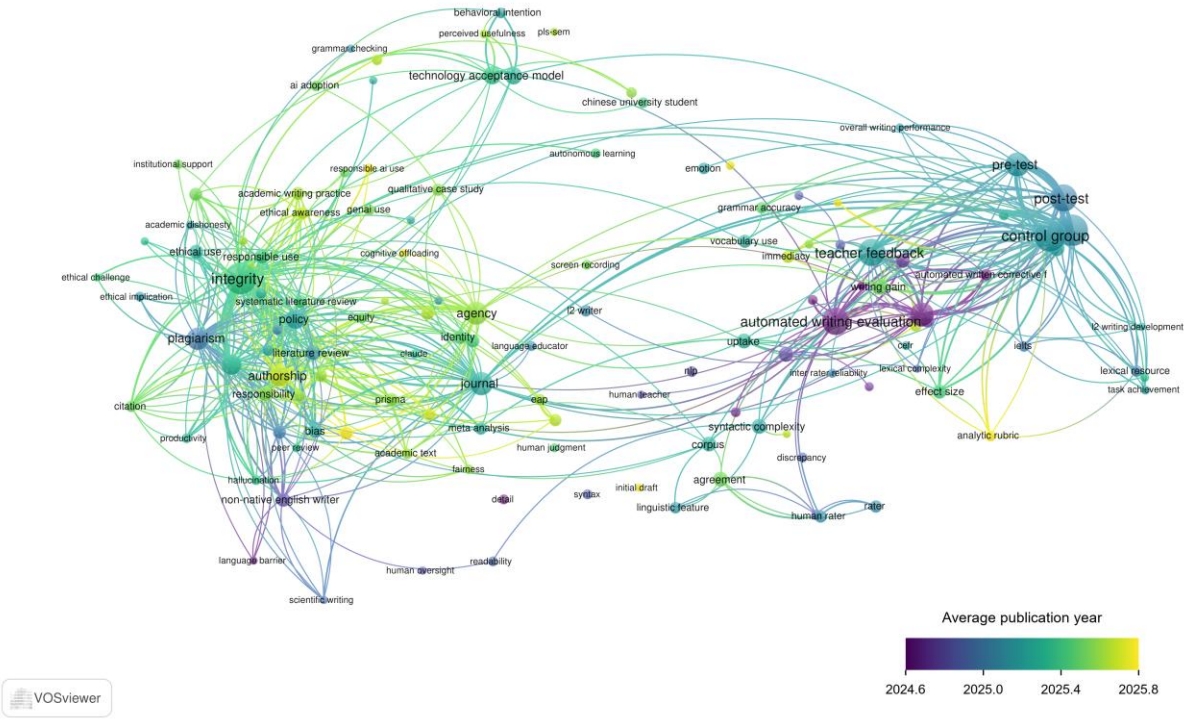


Figure 4. Overlay visualisation of the term map coloured by average publication year of the documents in which each term occurs (VOSviewer).

4.4 Intellectual base: co-citation clusters

The co-citation map of the 200 most co-cited references (Fig. 5) has three clusters (Table 6). K1, the largest (74 references; median year 2023), is the ChatGPT opportunity-and-risk wave: commentaries and early empirical studies on ChatGPT in L2 writing (Barrot, 2023; Su et al., 2023; Yan, 2023; Kohnke et al., 2023), efficacy studies (Song and Song, 2023; Marzuki et al., 2023; Mahapatra, 2024), general education debates (Kasneci et al., 2023; Cotton et al., 2023), the technology acceptance model (Davis, 1989) and the thematic analysis method (Braun and Clarke, 2006). K2 (63; median 2018) is the AWE and corrective-feedback foundation: engagement with automated feedback (Zhang and Hyland, 2018; Koltovskaia, 2020; Ranalli, 2018; Ranalli, 2021; Zhang, 2020), AWE effects (Link et al., 2020; Li et al., 2015; Barrot, 2021; Stevenson and Phakiti, 2014) and feedback theory (Hattie and Timperley, 2007; Hyland and Hyland, 2006; Hyland and Hyland, 2019). K3 (63; median 2024) is the GenAI feedback and assessment front: the quality of ChatGPT feedback (Guo and Wang, 2023; Escalante et al., 2023; Steiss et al., 2024), automated essay scoring with LLMs (Mizumoto and Eguchi, 2023), the affordances and contradictions of AI text for L2 writers (Warschauer et al., 2023), learner engagement and critical AI literacy (Teng, 2024;

Koltovskaia et al., 2024; Wang and Wang, 2025), feedback literacy (Carless and Boud, 2018) and the cognitive process model of writing (Flower and Hayes, 1981).

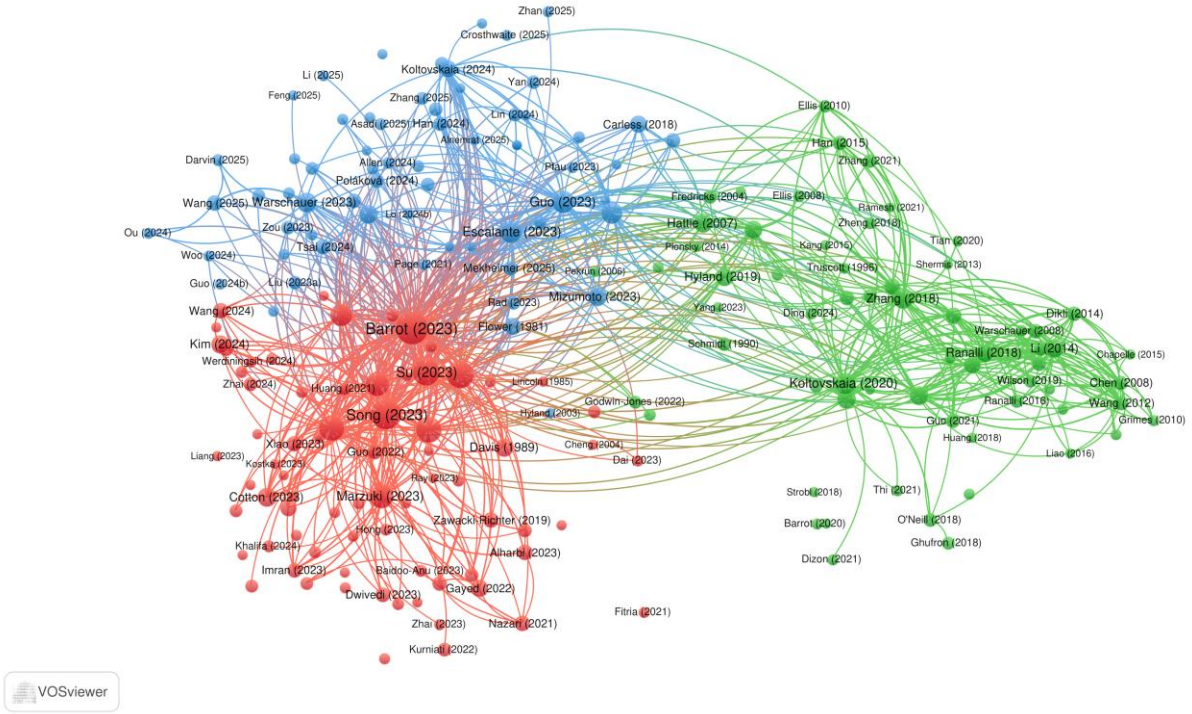


Figure 5. Reference co-citation network (VOSviewer; 200 references cited by ≥ 30 corpus documents). Node size: number of citing documents; colour: cluster (K1 red, K2 blue, K3 green).

Table 6. Co-citation clusters (intellectual base).

Cluster (refs; median year)	**Most co-cited references (citing documents)**	**Intellectual focus**
K1 ChatGPT opportunity–risk wave (74; 2023)	Barrot 2023 (348); Song and Song 2023 (297); Su et al. 2023 (240); Braun and Clarke 2006 (238); Yan 2023 (217); Kasneci et al. 2023 (207); Kohnke et al. 2023 (189); Marzuki et al. 2023 (178)	Early commentaries, perceptions and efficacy studies of ChatGPT; integrity debate; acceptance
K2 AWE and feedback foundations (63; 2018)	Koltovskaia 2020 (139); Zhang and Hyland 2018 (133); Link et al. 2020 (112); Hattie and Timperley 2007 (110); Barrot 2021 (109); Li et al. 2015 (107); Hyland and Hyland 2006 (105); Ranalli 2018 (102)	Engagement with automated and teacher feedback; AWE effects and limits
K3 GenAI feedback and assessment front (63; 2024)	Guo and Wang 2023 (153); Escalante et al. 2023 (152); Steiss et al. 2024 (150); Warschauer et al. 2023 (124); Mizumoto and Eguchi 2023 (116); Teng 2024 (114); Flower and Hayes 1981 (94); Carless and Boud 2018 (92)	Quality of ChatGPT feedback; LLM scoring; engagement and critical AI literacy

Note. Years are years of first publication; where OpenAlex and Crossref differ by one year (e.g., Li et al., 2015, labelled 2014 in Fig. 5), the reference list follows Crossref.

The co-citation structure shows that GenAI research has partly re-connected with its AWE ancestry: 28.2% of K2's co-citation strength links to K3, and K3 links more to K1 (43.7%)

than to itself (29.4%). Notably absent from the core of the intellectual base is the detection-fairness literature. The study that established bias of GPT detectors against non-native writers (Liang et al., 2023) was co-cited by only 32 of the 2,612 documents (1.2%), barely above the threshold for inclusion in the map, and no detection study ranked among the most co-cited works. The field's intellectual memory contains its feedback ancestors but not the assessment research most relevant to the writers it studies.

4.5 Influential studies

The most cited documents in the corpus (Table 7) are dominated by 2023–2024 studies of ChatGPT and by earlier AWE engagement studies. Their designs are telling: of the ten most cited, three are commentaries or exploratory studies, two are classroom interventions, two are qualitative perception or engagement studies, one compares ChatGPT and teacher feedback, one is an LLM scoring study and only one is a randomised trial. The heavy citation of early opinion and exploratory work reflects the speed of the field; it also means that claims established before robust evidence existed continue to anchor the literature.

Table 7. Ten most cited documents in the corpus (OpenAlex citations at retrieval).

Study	Focus and design	Cites
Song and Song (2023)	ChatGPT-assisted instruction; experimental; writing skills and motivation	1,092
Barrot (2023)	Pitfalls and potentials of ChatGPT for L2 writing; commentary	1,027
Yan (2023)	ChatGPT in an L2 writing practicum; exploratory	887
Marzuki et al. (2023)	AI writing tools and content/organisation; teacher perspectives (qualitative)	809
Mizumoto and Eguchi (2023)	LLM-based automated essay scoring; computational	752
Mahapatra (2024)	ChatGPT as formative feedback for ESL academic writing; mixed-methods intervention	678
Su et al. (2023)	ChatGPT in argumentative writing classrooms; exploratory	678
Nazari et al. (2021)	AI-powered writing assistant; randomised controlled trial	554
Guo and Wang (2023)	ChatGPT versus teacher feedback; comparative	535
Koltovskaia (2020)	Engagement with Grammarly feedback; case study	446

Note. Full list of 25 in Supplementary Table S2 (top_cited_corpus.csv).

4.6 Evidence profile and gaps (RQ3–RQ4)

Abstract-level coding characterises the evidence base (Table 8). Qualitative (39.5%) and survey or structural-equation designs (33.7%) together outnumber experimental and quasi-experimental studies (20.1%), and affective outcomes (confidence, anxiety, motivation, engagement) were mentioned in 42.4% of abstracts. Explicit theoretical framing was reported in only 24.2% of abstracts, although the share rose from 15.1% before 2023 to 31.1% in 2026. Teachers were the focal population in 9.3% of records, and ESP contexts in only 3.6% (87 records), of which 12 concerned ESP teachers and 18 used experimental designs.

Table 8. Indicators from abstract-level coding (records with abstracts, n = 2,448).

Indicator	n	%
Experimental or quasi-experimental design	492	20.1
Survey or structural equation modelling	824	33.7
Qualitative design (interviews, case study, thematic analysis)	968	39.5
Review or bibliometric study	211	8.6
Textual or process data (corpus, screen recording, keystroke)	210	8.6
GenAI / LLM	1,485	60.7
AWE or automated feedback	369	15.1
Grammarly	340	13.9
QuillBot	145	5.9
Machine translation	79	3.2
ESP or professional context	87	3.6
EAP or research writing	224	9.2
School context	128	5.2
Teachers as focal population	228	9.3
Affective outcomes	1,038	42.4
Critical thinking or argumentation	455	18.6
Agency, authorship, voice or identity	323	13.2
Integrity or plagiarism	399	16.3
AI-text detection	61	2.5
Bias or false positives concerning non-native writers	15	0.6
Delayed post-test or longitudinal design	188	7.7
Transfer to unaided writing	117	4.8
Explicit theoretical framing	593	24.2

Note. Dictionary-based coding of titles and abstracts; categories are not mutually exclusive. Dictionaries: summarise.py.

The profile of the 492 intervention studies is the most consequential result for claims about learning (Fig. 7b). Only 48 (9.8%) mentioned a delayed post-test or longitudinal measurement, 28 (5.7%) examined transfer to writing without AI support, 80 (16.3%) measured critical thinking or argumentation, 24 (4.9%) addressed agency or authorship, 21 (4.3%) addressed integrity and 18 (3.7%) were situated in ESP or professional contexts. AI-

text detection appeared in 61 records overall, of which only 10 also discussed bias or false positives affecting non-native writers and only six involved experimental designs or AWE, that is, the pedagogical settings in which L2 writers' texts are actually transformed by AI.

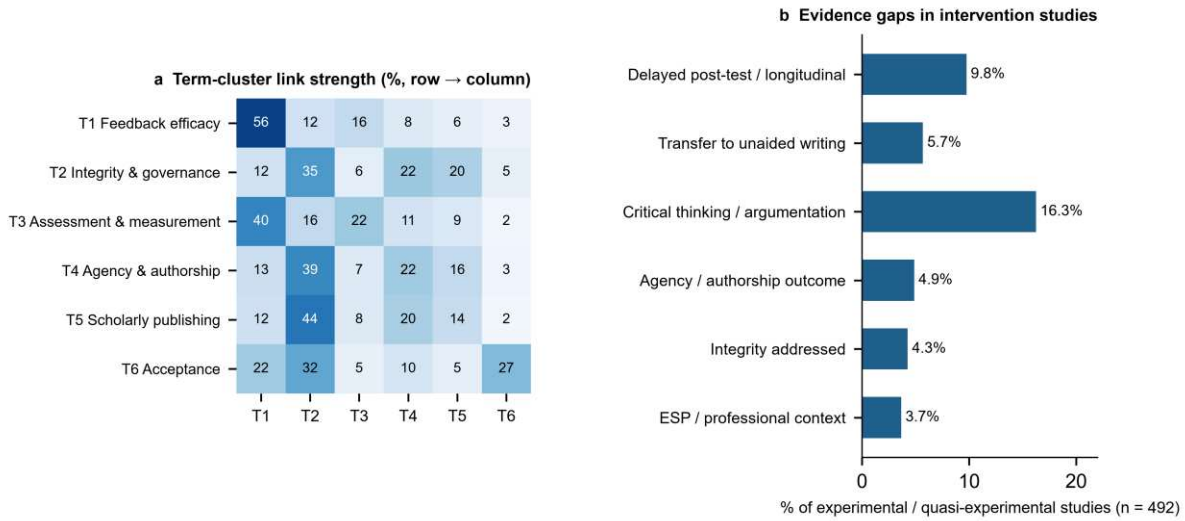


Figure 7. (a) Share of each term cluster's link strength directed to every cluster (rows sum to 100%). (b) Share of experimental and quasi-experimental studies (n = 492) whose abstracts report selected design features or outcomes.

5. Critical synthesis of the evidence

This section interprets the clusters in the light of the studies that define them, moving from what the evidence supports with reasonable confidence to what it leaves open.

5.1 Text-level gains are consistent but proximal (T1, K1)

Across contexts and tools, intervention and perception studies converge on improvements in grammatical accuracy, vocabulary, organisation and coherence. Experimental studies in China, Yemen, India and Indonesia report higher writing scores for ChatGPT-assisted groups (Song and Song, 2023; Alkamel and Alwagieh, 2024; Mahapatra, 2024; Amanda et al., 2026); a randomised trial found that an AI writing assistant improved learners' engagement and self-efficacy (Nazari et al., 2021); studies of QuillBot report gains in writing performance and confidence and reduced apprehension (El-Garawany, 2024; Rozi et al., 2026); and learners value machine feedback from Grammarly alongside teacher feedback (Kawashima, 2023). Similar perceived benefits are reported for non-native medical students writing research papers (Li et al., 2024) and for learners of Chinese as a second language (Chen and Gong, 2025), which suggests that the pattern is not specific to English. Teachers and students in Indonesia, Saudi Arabia and elsewhere perceive AI tools as improving content,

organisation and efficiency (Marzuki et al., 2023; Selim, 2024; Ozfidan et al., 2024; Al-Sofi, 2024), and qualitative work suggests that AI-mediated writing can help learners notice weaknesses and revise (Fathi and Rahimi, 2024). These benefits match the problems most often reported for EFL academic writers (Raheem and Ghafar, 2025).

Three limitations qualify this convergence. First, the outcome is usually a post-test written shortly after, and often in the same conditions as, the intervention; only a small minority of studies measure retention or performance without AI (Section 4.6). The evidence therefore establishes mediated accuracy more firmly than learning. Earlier AWE research reached the same conclusion: computer-generated feedback improved the quality of revised texts, but evidence of transfer to general writing proficiency was limited (Stevenson and Phakiti, 2014). Second, many studies measure perceptions rather than performance. In the business English study, for instance, all 50 students agreed that QuillBot improved their grammar, but no text-based measure was collected (Rozi et al., 2026). Perceived benefit and learning are distinct constructs, and the high share of survey and qualitative designs means that much of the corpus speaks to the former. Third, comparison conditions are often weak. Conventional instruction without feedback is not an equivalent control for an intervention that adds immediate, individual feedback; the relevant counterfactual is high-quality human or peer feedback of similar intensity. Studies that use such comparisons yield more nuanced results (Section 5.2).

5.2 From AWE to GenAI feedback: engagement is the mechanism (T1, K2, K3)

The AWE literature (K2) confirmed that it is the interaction between learners and automated feedback that determines the value of the automated feedback. The case studies revealed that engagement is a complex and uneven phenomenon, with some learners deeply processing feedback and others simply taking it in or rejecting it without any cognitive processing (Zhang and Hyland, 2018; Koltovskaia, 2020; Zhang, 2020). The effectiveness of feedback determined if learners could use it (Ranalli, 2018; Ranalli, 2021) and teacher integration was important: AWE was not an alternative or replacement for teacher feedback, and how teachers position AWE influenced its benefits for learners (Li et al., 2015; Link et al., 2020). These are exactly the conditions that the feedback-engagement lens (Section 2.2) foresees.

This picture is mostly supported by research on GenAI feedback, known as K3. ChatGPT generates more feedback than teachers, covering content, organisation and language (Guo and Wang, 2023), and the quality of the feedback is comparable to or approaches that of

trained human raters (Steiss et al., 2024), but is not, in general, equal to that of human raters. When compared to human tutoring, student learning outcomes were comparable and student preferences were split when using AI-generated feedback (Escalante et al., 2023). The effects of LLM-generated feedback on revision performance and motivation have been demonstrated in secondary EFL classrooms (Meyer et al., 2024), and the use of ChatGPT in teacher feedback has been examined to incorporate the efficiency of AI with teacher judgement (Han and Li, 2024). The usage of ChatGPT for rewriting L2 writing by graduate writers in terms of their behavioural, cognitive and affective engagement in the process has been quite diverse, ranging from critical evaluation to full acceptance (Koltovskaia et al., 2024). The agenda is that the nature of feedback is transformed by the introduction of GenAI, but still engagement and feedback literacy (Carless and Boud, 2018; Rad et al., 2023) are essential to learning. If a study considers AI as a treatment, instead of as a tool that influences the treatment through engagement, then it risks assigning to the tool that which is due to the engagement.

5.3 AI as assessor and the linguistic profile of AI-assisted text (T3)

The field of AI as a rater and the linguistic characteristics of AI-generated texts is a unique strand. LLMs can score L2 essays with a useful level of accuracy and reliability, which improves when linguistic features are added (Mizumoto and Eguchi, 2023). The results of corpus studies indicate that the use of lexis and collocation of the AI-generated academic essays differs from that of L2 learner essays (Zhang and Crosthwaite, 2025). The results of this study are two-pronged. The features that enable machines to score proficiency are the features that differentiate more and less proficient L2 writers (Crossley and McNamara, 2010), and the features that AI assistance changes. Because fairness is in the assessment cluster but not connected with the non-native writer concept in T5 (Section 4.2), this knowledge has not yet been used with the writers whose texts are being judged.

5.4 Agency, authorship and critical AI literacy: the emerging front (T4)

The latest cluster takes a different approach to the question, shifting its focus from outcomes to agency. Initial analysis of ChatGPT highlighted its potential for generating models and feedback but also suggested that it might have a negative impact on learning and confuse authorship (Barrot, 2023; Warschauer et al., 2023). Empirical research in recent years investigates how L2 writers actually work with AI to brainstorm, outline, revise, edit, source texts, and more, along with aspects of critical AI literacy like critical awareness, positionality, interaction strategies, and judging AI affordances (Wang and Wang, 2025). Studies of

human–AI collaboration show that interaction patterns differ and that more iterative, critically engaged patterns are associated with better writing (Nguyen et al., 2024), whereas learners themselves describe AI as a companion whose use requires judgement (Teng, 2024; Kim et al., 2024; Chan and Hu, 2023).

This front is directly addressing the sociocultural difference between scaffolding and substitution. The distinction is of critical importance because evidence on cognitive offloading suggests that the heavier the reliance on AI tools, the lower the critical-thinking scores, and that use of AI tools is linked to a decrease in neural engagement, and thus sense of ownership, for essay writers, which was reported in a preprint study and was not carried out with L2 writers (Gerlich, 2025; Kosmyna et al., 2025). The L2 academic writing context may be more demanding regarding the cost of substitution since the writer is developing two voices; a language and a disciplinary voice. However, of these, only 4.9% dealt with agency or authorship (Section 4.6). The emerging front is primarily qualitative so far, and has yet to be tied to intervention designs to determine if there is a moderating effect of critical AI literacy on learning gains.

5.5 Integrity, detection and the fairness problem (T2, T5)

The integrity cluster is large but mostly normative. It contains guidelines, policies, ethical frameworks and reviews concerned with plagiarism, responsible use and over-reliance (Cotton et al., 2023; Dergaa et al., 2023; Deep and Chen, 2025; Ofem et al., 2024), and it reflects real institutional uncertainty. ESP teachers, for example, identified unclear institutional policy as a key factor in their ambivalence towards GenAI (Alrashed et al., 2026). However, the cluster contains little empirical work on how integrity judgements are actually made about L2 writers' texts.

That empirical work exists mainly outside the core of the field. Detectors have been shown to be biased against non-native writers (Liang et al., 2023), inaccurate and unreliable (Weber-Wulff et al., 2023), and sensitive to genre, length and hybrid authorship (Fraser et al., 2025; Hadra et al., 2026). Within L2 writing, RoBERTa-based classifiers detected ChatGPT essays with inconsistent accuracy (Ibrahim, 2023), and novice teachers varied considerably in their ability to identify AI-generated EFL texts (De Wilde, 2024). The most direct evidence comes from the paired-document study of professional editing: holding content and authorship constant, linguistic revision alone shifted detector scores, in opposite directions for different detectors and in proportion to the amount of editing, and false-positive rates on unedited

human text ranged from 0% to 100% (Park et al., 2026). Because that study is a preprint, its estimates require peer-reviewed confirmation, but its design isolates style from authorship more cleanly than population comparisons can.

These findings matter for the pedagogical literature because the improvements reported in T1 are changes in style. If AI feedback makes an L2 writer's text more fluent, cohesive and lexically sophisticated, and if detectors respond to such features, then the same intervention that raises a writing score may raise the probability that the writer's later, independent work is treated as suspect, whether or not AI was used. Detection research in the corpus is rare (2.5%), and only six records connected detection to the experimental or AWE settings in which this would occur. Calls to move from detection-based policing to assessment redesign, for instance by specifying permitted levels of AI use or changing the structure of tasks, address this risk from the assessment side (Perkins et al., 2024; Corbin et al., 2025), but they have not yet been evaluated with L2 writers.

5.6 ESP, teachers and the limits of acceptance models (T6)

ESP is strikingly under-represented (3.6% of records) relative to its importance. ESP writing is genre-specific and often high-stakes, and AI tools may help with professional conventions such as register and formulaic expressions (Rozi et al., 2026); yet the genre knowledge that ESP pedagogy aims to develop is precisely what generic AI output may flatten (Hyland, 2025). Teacher research shows that ESP teachers' responses are shaped by affective dispositions and by social and institutional negotiation (Alrashed et al., 2026; Verano-Tacoronte et al., 2025), which acceptance models capture only partially. The technology acceptance cluster (T6) is small and comparatively isolated; it explains intention to use AI with perceived usefulness and ease of use (Davis, 1989; Venkatesh et al., 2003), but intention says little about whether use is pedagogically productive or fairly assessed.

6. Discussion

6.1 The mediated-accuracy paradox

Integrating the bibliometric and critical analyses yields a coherent account of the field. Research on AI in EFL/ESP academic writing has grown explosively and shifted from AWE to GenAI. It has produced consistent evidence of short-term text-level gains and a rich understanding of engagement, but it has rarely tested durable learning or transfer, and it has not connected its pedagogical findings with the assessment regimes that judge AI-assisted

texts. The science map shows this disconnection structurally (16.1% of link strength between the pedagogical and integrity–authorship blocks); the co-citation analysis shows that detection-fairness research is marginal to the field's intellectual base; and the coding shows that interventions seldom measure retention, transfer, agency or integrity.

We call the resulting tension the mediated-accuracy paradox: the linguistic polish that AI makes available to L2 writers, which the pedagogical literature measures as a gain, is also a feature on which style-sensitive integrity judgements, whether made by detectors or by teachers, may act. The paradox is not that AI help is bad, but that two research communities evaluate the same textual change against different criteria without reference to each other. Its consequences fall disproportionately on L2 writers, for whom the gap between unaided and AI-mediated prose is largest (Liang et al., 2023; Park et al., 2026). This reframing extends earlier warnings about the affordances and contradictions of AI text for L2 writers (Warschauer et al., 2023) by locating the contradiction in the relation between pedagogy and assessment and by showing that the literature's structure reproduces it.

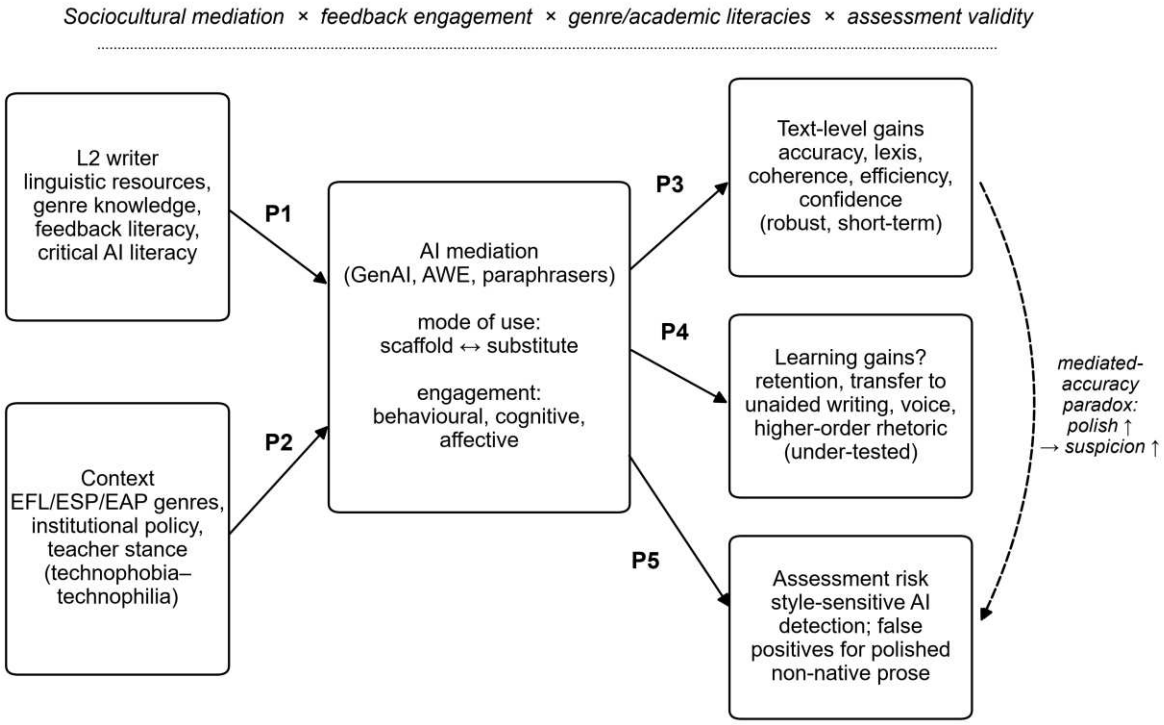


Figure 8. Integrative framework of AI-mediated L2 academic writing. P1–P5 denote the propositions in Section 6.2; the dashed arc represents the mediated-accuracy paradox.

6.2 Integrative framework and propositions

Figure 8 integrates the four theoretical lenses into a framework that links the L2 writer and context to AI mediation, and AI mediation to three classes of outcome. Five propositions follow from the synthesis and are formulated so that they can be tested.

- P1. The effect of AI mediation on learning depends on writers' feedback literacy and critical AI literacy: writers who evaluate and selectively appropriate AI suggestions will show larger retained gains than writers who accept them wholesale (Carless and Boud, 2018; Wang and Wang, 2025; Koltovskaia et al., 2024).
- P2. Context shapes the mode of AI use: genre demands (ESP, EAP), institutional policy and teacher stance determine whether AI functions as a scaffold or a substitute (Alrashed et al., 2026; Hyland, 2025).
- P3. AI mediation reliably improves proximal, text-level outcomes (accuracy, lexis, coherence, efficiency, confidence) regardless of mode of use (Song and Song, 2023; Mahapatra, 2024; Amanda et al., 2026).
- P4. Durable learning gains, namely retention, transfer to unaided writing, voice and higher-order rhetoric, occur only when AI is used as a contingent and fading scaffold, and are smaller than proximal gains (Wood et al., 1976; Stevenson and Phakiti, 2014).
- P5. The stylistic changes produced by AI mediation alter the probability that L2 writers' texts are judged AI-generated, and this effect is larger for writers whose unaided and mediated prose differ most (Park et al., 2026; Liang et al., 2023).

6.3 Theoretical implications

The review has three theoretical implications. First, the distinction between mediated performance and internalised competence, basic to sociocultural theory, should be built into outcome measurement: AI-assisted and unaided performance are different dependent variables, and conflating them inflates estimates of learning. Second, feedback-engagement theory transfers well from AWE to GenAI and offers a mechanism that the many tool-comparison studies lack; research should model engagement and feedback literacy as mediators and moderators rather than describing them separately. Third, assessment validity must be treated as part of the theory of AI-mediated writing, not as an external ethical issue. Once the inference from text features to authorship is recognised as fallible and style-dependent, fairness becomes an outcome of AI-mediated pedagogy that can be studied empirically.

6.4 Practical implications

For teachers, the evidence supports using AI feedback as one source among several, embedded in tasks that require learners to evaluate, accept or reject suggestions and to explain their choices, and fading support over time. Process evidence such as drafts, revision histories and reflective commentaries provides a more valid basis for judging learning, and authorship, than the final text. For ESP programmes, AI tools should be aligned with explicit genre instruction so that learners appropriate disciplinary conventions rather than generic fluency (Hyland, 2025; Rozi et al., 2026). The detection evidence suggests that for institutions, the scores of the detectors are not evidence of misconduct to be relied upon alone or primarily, especially for L2 writers (Weber-Wulff et al., 2023; Liang et al., 2023; Giray, 2024); clear policies that spell out permitted uses of AI by task, and assessment designs that make the writing process visible, address the policy vacuum that teachers report (Perkins et al., 2024; Corbin et al., 2025; Alrashed et al., 2026). In the field of applied linguistics, the design profile reported here recommends that journal editors and reviewers require intervention studies to report on delayed and unaided measures and to justify the conditions that they compare.

6.5 Research agenda

Table 9 translates the gaps into a research agenda linked to the propositions.

Table 9. Research agenda derived from the integrated analysis.

Gap (evidence)	Research priority	Suggested design
Durable learning rarely tested (9.8% delayed; 5.7% transfer)	Separate mediated from unaided performance; test retention (P3, P4)	Randomised or quasi-experimental designs with delayed, AI-free post-tests and active, intensity-matched comparison feedback
Engagement studied apart from outcomes	Model feedback literacy and critical AI literacy as moderators (P1)	Process-tracing (screen recording, logs, keystroke) combined with outcome measures
Pedagogy–integrity disconnection (16.1% cross-block links)	Test whether AI-mediated gains change detection and teacher judgements (P5)	Paired-text designs: unaided vs AI-revised texts by the same L2 writers scored by detectors and blinded teachers
Detection fairness marginal (2.5%; 0.6% on bias)	Evaluate detector and human-judgement validity for L2 populations, genres and proficiency levels	Validity studies with ground-truth authorship; differential false-positive analysis
ESP under-represented (3.6%)	Genre-specific effects and genre appropriation (P2)	Genre-based interventions in professional domains; move analysis of AI-assisted texts
Agency front mostly qualitative	Link voice, ownership and offloading to learning outcomes	Mixed-methods longitudinal cohorts; validated measures of authorial voice
Theory reported in 24.2% of abstracts	Theory-driven hypotheses rather than tool comparisons	Pre-registered studies testing propositions P1–P5
Volume–visibility asymmetry (e.g., Indonesia)	Multi-site, cross-national collaboration	Coordinated replications across EFL contexts with shared instruments

7. Limitations

There are a number of constraints to be kept in mind. To begin with, the bibliometric analysis is based on a single database. OpenAlex is selected for its openness and its coverage of regional journals, but some errors in its metadata and abstracts are missing from the search results, which are mitigated but not eliminated by the supplementary journal-scope search, the eligibility rule for title-only search and language checks, and can vary from analyses of Web of Science or Scopus. Secondly, rule-based screening replaces the judgement with transparency and scale from manual screening. The precision audit (94.9% fully relevant) is reassuring but one reviewer has performed the precision audit and recall cannot be estimated exactly. Third, abstract-level coding is used in order to detect what is reported in the abstracts; features not reported in abstracts, including delayed post-tests, will be under-counted, so the gap indicators should be interpreted as lower bounds on reporting, rather than being treated as prevalence values. Fourth, the essence of citation-based analyses is to privilege older texts, and 79.5% of the corpus is dated from 2025–2026, so the intellectual base as described in the co-citation map reflects this one and not the eventual influence of new texts. Fifth, the close reading was not comprehensive, and there was no formal risk-of-bias appraisal of individual studies. Finally, one of the five seed articles (Park et al., 2026) is a preprint, and its findings, like those of the EEG preprint cited in Section 5.4, should be treated as provisional.

8. Conclusion

This review combined VOSviewer science mapping of 2,612 documents with a critical synthesis to examine research on AI in EFL and ESP academic writing from 2019 to 2026. The field has grown exponentially and pivoted from AWE to generative AI. Its strongest evidence concerns short-term improvements in the accuracy, organisation and fluency of L2 texts and the role of learner engagement in realising them. Its weakest concerns whether those improvements are retained and transfer to unaided writing, how they bear on writers' agency and voice, and how they interact with the integrity judgements that institutions make. The pedagogical and integrity strands of the literature are weakly connected, and research on the fairness of AI detection for non-native writers remains marginal to the field's intellectual base. The mediated-accuracy paradox names the consequence: the polish that makes AI

useful to L2 writers can also make their writing suspect. Resolving it requires research that measures learning rather than mediated performance, models engagement and critical AI literacy as mechanisms, attends to ESP genres and evaluates assessment practices for L2 writers. The framework and agenda proposed here are offered as a basis for that work.

Declarations

Funding. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests. The authors declare no competing interests.

Ethics approval. Not applicable. The study analysed published bibliographic records and did not involve human participants.

Data and code availability. All data (OpenAlex records, CC0), screening logs, thesaurus files, VOSviewer map and network files, coding dictionaries and Python scripts are provided in Supplementary File S1 and will be deposited in an open repository upon publication.

Use of AI tools. Generative AI was used to assist with code for data processing and with language editing. The authors designed the study, checked every analysis against the underlying files, verified all references against Crossref and take full responsibility for the content.

Author contributions. [To be completed by the authors using the CRediT taxonomy.]

Supplementary material

Supplementary File S1 (reproducibility package): search strings, raw and screened OpenAlex records, screening and exclusion logs, precision-audit sample, thesaurus files, VOSviewer map and network files, coding dictionaries, tables and Python scripts. Figure S1: term density visualisation. Figure S2: co-citation density visualisation. Table S1: full term map with clusters, occurrences and average publication years. Table S2: 25 most cited documents.

References

Akram, M.F., Ali, M.R. and Kadirova, Z. (2023) 'Role of consumer digital literacy in consumer behavioral intentions: Mediating effect of perceived ease of use, usefulness and customer engagement', *Multidisciplinary Journal of Educational Research*, 13(1), pp. 104–130.

Al-Sofi, B. (2024) 'Artificial intelligence-powered tools and academic writing: to use or not to use ChatGPT', *Saudi Journal of Language Studies*, 4(3), pp. 145–161. Available at: https://doi.org/10.1108/sjls-06-2024-0029.

Ali, O., Murray, P.A., Momin, M., Dwivedi, Y.K. and Malik, T. (2024) 'The effects of artificial intelligence applications in educational settings: Challenges and strategies', *Technological Forecasting and Social Change*, 199, 123076. Available at: https://doi.org/10.1016/j.techfore.2023.123076.

Alkamel, M.A.A. and Alwagieh, N.A.S. (2024) 'Utilizing an adaptable artificial intelligence writing tool (ChatGPT) to enhance academic writing skills among Yemeni university EFL students', *Social Sciences & Humanities Open*, 10, 101095. Available at: https://doi.org/10.1016/j.ssaho.2024.101095.

Alperin, J.P., Portenoy, J., Demes, K., Larivière, V. and Haustein, S. (2024) 'An analysis of the suitability of OpenAlex for bibliometric analyses', *arXiv* [Preprint], arXiv:2404.17663. Available at: https://doi.org/10.48550/arXiv.2404.17663.

Alrashed, M.A., Khalil, A.A., Ali El Deen, A.A.M.M., Naby, S.M.A., Wannas, A.S. and Mohamed, A.M. (2026) 'ESP Teachers' Perceptions of Generative AI in University ESP Writing Instruction between Technophobia and Technophilia', *World Journal of English Language*, 16(6), 298. Available at: https://doi.org/10.5430/wjel.v16n6p298.

Amanda, V., Darmayanti, Y., Zahra, R., Ningsih, A.R., Muafi, K. and Oktarisna, R. (2026) 'The Effect of ChatGPT-Assisted Writing on EFL Students' Academic Writing Skills', *Journal of English Language and Education*, 11(3), pp. 726–736. Available at: https://doi.org/10.31004/jele.v11i3.2477.

Banh, L. and Strobel, G. (2023) 'Generative artificial intelligence', *Electronic Markets*, 33(1), 63. Available at: https://doi.org/10.1007/s12525-023-00680-1.

Barrot, J.S. (2021) 'Using automated written corrective feedback in the writing classrooms: effects on L2 writing accuracy', *Computer Assisted Language Learning*, 36(4), pp. 584–607. Available at: https://doi.org/10.1080/09588221.2021.1936071.

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

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

Carless, D. and Boud, D. (2018) 'The development of student feedback literacy: enabling uptake of feedback', *Assessment & Evaluation in Higher Education*, 43(8), pp. 1315–1325. Available at: https://doi.org/10.1080/02602938.2018.1463354.

Chan, C.K.Y. and Hu, W. (2023) 'Students' voices on generative AI: perceptions, benefits, and challenges in higher education', *International Journal of Educational Technology in Higher Education*, 20(1), 43. Available at: https://doi.org/10.1186/s41239-023-00411-8.

Chen, C. and 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. Available at: https://doi.org/10.3390/educsci15020141.

Corbin, T., Dawson, P. and Liu, D. (2025) 'Talk is cheap: why structural assessment changes are needed for a time of GenAI', *Assessment & Evaluation in Higher Education*, 50(7), pp. 1087–1097. Available at: https://doi.org/10.1080/02602938.2025.2503964.

Cotton, D.R.E., Cotton, P.A. and Shipway, J.R. (2023) 'Chatting and cheating: Ensuring academic integrity in the era of ChatGPT', *Innovations in Education and Teaching International*, 61(2), pp. 228–239. Available at: https://doi.org/10.1080/14703297.2023.2190148.

Crossley, S.A. and McNamara, D.S. (2010) 'Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication', *Journal of Research in Reading*, 35(2), pp. 115–135. Available at: https://doi.org/10.1111/j.1467-9817.2010.01449.x.

Culbert, J.H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P. and Mayr, P. (2025) 'Reference coverage analysis of OpenAlex compared to Web of Science and Scopus', *Scientometrics*, 130(4), pp. 2475–2492. Available at: https://doi.org/10.1007/s11192-025-05293-3.

Ali, M., Kadirova, Z. and Ismoyilov, B., 2024, June. Practicing green HRM to achieve environmental sustainability. In Environment. Technology. Resources. Proceedings of the International Scientific and Practical Conference (Vol. 1, pp. 24-33).

https://doi.org/10.17770/etr2024vol1.7961

Ali, M., Peiseniece, L. and Mikelsone, E., 2025, June. Through the green work commitment, a literature review examines the relationship between green HRM and environmental sustainability. In Environment. Technology. Resources. Proceedings of the International Scientific and Practical Conference (Vol. 1, pp. 31-43).

https://doi.org/10.17770/etr2025vol1.8635

Ali, M., Peiseniece, L. and Mikelsone, E., 2025, June. The mediating role of green work commitment in the relationship between green HR practices, AI elements, and environmental sustainability: a quantitative analysis. In ENVIRONMENT. TECHNOLOGY. RESOURCES. Proceedings of the International Scientific and Practical Conference (Vol. 2, pp. 15-22).

https://doi.org/10.17770/etr2025vol2.8578

Shah, A.S.A., Ali, M.R., Vokhidova, M., Xusanov, C. and Musaeva, R., 2025. Charting the Future: Analyzing Innovative Strategies in Strategic Management for Sustainable Performance Excellence and Business Leadership in the Telecom Industry. Business Ethics and Leadership, 9(4), pp.195-208.

https://doi.org/10.61093/bel.9(4).195-208.2025

Khan, S., Ali, M., Akram, M., Azizova, N., Vokhidova, M. and Isamukhamedova, G., 2025, June. Exploring the impact of digital transformation on business model innovation in the retail sector. In Environment. Technology. Resources. Proceedings of the International Scientific and Practical Conference (Vol. 1, pp. 281-288).

https://doi.org/10.17770/etr2025vol1.8640

Qureshi, A., Manzoor, A., Khan, S. and Ali, M.R., 2024, December. Utilizing deep learning methods for stress detection in social media platform. In Proceedings of the 8th International Conference on Future Networks & Distributed Systems (pp. 660-664).

https://doi.org/10.1145/3726122.3726216

Ali, M.R., Peiseniece, L., Mikelsone, E., Shah, A.S.A. and Kadirova, Z., 2026. Co-Creating Sustainable Innovation From Within: Green HRM Practices Enabling EmployeeGreen Behavior in KIBS Ecosystems Across Pakistan, China, And Uzbekistan. Journal of Management Practices, Humanities and Social Sciences, 10(2), pp.263-277.

https://doi.org/10.33152/jmphss-10.2.7

Bakhronkulova, L., Ali, M., Khabirova, Z., Azimjon, A., Zebo, A. and Muslima, A., 2025, June. Artificial Intelligence in Cybersecurity: Threats, Defenses, and Future Directions. In ENVIRONMENT. TECHNOLOGY. RESOURCES. Proceedings of the International Scientific and Practical Conference (Vol. 2, pp. 23-30).

https://doi.org/10.17770/etr2025vol2.8592

Ali M.R., Peiseniece L. & Mikelsone E. Fostering environmental sustainability through green HRM green leadership and work commitment. Discov Sustain (2026). https://doi.org/10.1007/s43621-026-04468-z

Davis, F.D. (1989) 'Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology', *MIS Quarterly*, 13(3), pp. 319–340. Available at: https://doi.org/10.2307/249008.

De Wilde, V. (2024) 'Can novice teachers detect AI-generated texts in EFL writing?', *ELT Journal*, 78(4), pp. 414–422. Available at: https://doi.org/10.1093/elt/ccae031.

Deep, P.D. and Chen, Y. (2025) 'The Role of AI in Academic Writing: Impacts on Writing Skills, Critical Thinking, and Integrity in Higher Education', *Societies*, 15(9), 247. Available at: https://doi.org/10.3390/soc15090247.

Dergaa, I., Chamari, K., Zmijewski, P. and Ben Saad, H. (2023) 'From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing', *Biology of Sport*, 40(2), pp. 615–622. Available at: https://doi.org/10.5114/biolsport.2023.125623.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N. and Lim, W.M. (2021) 'How to conduct a bibliometric analysis: An overview and guidelines', *Journal of Business Research*, 133, pp. 285–296. Available at: https://doi.org/10.1016/j.jbusres.2021.04.070.

El-Garawany, M.S.M. (2024) 'The Effects of a QuillBot-Based Intervention on English Language Majors' EFL Writing Performance, Apprehension, and Self-Efficacy', *Language Teaching Research Quarterly*, 43, pp. 167–189. Available at: https://doi.org/10.32038/ltrq.2024.43.10.

Ellis, R. (2010) 'Epilogue: A framework for investigating oral and written corrective feedback', *Studies in Second Language Acquisition*, 32(2), pp. 335–349. Available at: https://doi.org/10.1017/s0272263109990544.

Escalante, J., Pack, A. and 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), 57. Available at: https://doi.org/10.1186/s41239-023-00425-2.

Ali, M.R. and Iqbal, S.M.J., 2022. Examining the mediating influence of green behavior on the impact of green HR practices on environmental sustainability. iRASD Journal of Management, 4(3), pp.479-494.

Muhammad, R.A., Bakhtiyor, I., Mahliyo, M. and Muhammad, F.A., 2023. Role of Servant Leadership in Work Performance: Identifing the Mediating Relation with Employee Ambidexterity and Moderating with Competitive Orientation. Oeconomia Copernicana, 14(1), pp.181-211.

Ali, M.R., Akram, M.F., Urkinbayev, T.A., Tashtayeva, S.K. and Aripdjanov, G.T., 2023. The Pathways to Achieve Green Economy and The Process of Sustainable Development Goals. SGS-Engineering & Sciences, 2(02).

Kadirovich, X.C., Akmalevna, N.F., Shahzad, A., Ali, M.R., Kadirova, Z. and Akram, M.F., 2026. Governance Trust and Investment Behaviour as Pathways to Sustainable Hotel Development in Uzbekistan. https://doi.org/10.21203/rs.3.rs-11137556/v1

Ali, M.R., Asghar, T., Kadirova, Z., Khasanovna, V.M. and Akram, M.F., 2026. Artificial Intelligence in Green Human Resource Management: Examining the Mediating Role of Employee Engagement in Attaining Organizational Sustainability.

https://doi.org/10.21203/rs.3.rs-11412497/v1

Ali, M.R., 2026. The Incorporation and Transformation of Spiritual Education within the State Ideology of the People's Republic of China. https://doi.org/10.21203/rs.3.rs-11380596/v1

Kadirovich, X.C., Shah, A.S.A., Ali, M.R. and Khasanovna, V.M., 2026. The Moderating Role of Staff Training and Mediating Role of Venue Reputation in Accessible MICE Tourism: Effects on Business Event Satisfaction. https://doi.org/10.21203/rs.3.rs-11106800/v1

Fathi, J. and Rahimi, M. (2024) 'Utilising artificial intelligence-enhanced writing mediation to develop academic writing skills in EFL learners: a qualitative study', *Computer Assisted Language Learning*, 39(1-2), pp. 263–302. Available at: https://doi.org/10.1080/09588221.2024.2374772.

Feuerriegel, S., Hartmann, J., Janiesch, C. and Zschech, P. (2023) 'Generative AI', *Business & Information Systems Engineering*, 66(1), pp. 111–126. Available at: https://doi.org/10.1007/s12599-023-00834-7.

Flower, L. and Hayes, J.R. (1981) 'A Cognitive Process Theory of Writing', *College Composition & Communication*, 32(4), pp. 365–387. Available at: https://doi.org/10.58680/ccc198115885.

Fraser, K.C., Dawkins, H. and Kiritchenko, S. (2025) 'Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods', *Journal of Artificial Intelligence Research*, 82, pp. 2233–2278. Available at: https://doi.org/10.1613/jair.1.16665.

Gerlich, M. (2025) 'AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking', *Societies*, 15(1), 6. Available at: https://doi.org/10.3390/soc15010006.

Giray, L. (2024) 'The Problem with False Positives: AI Detection Unfairly Accuses Scholars of AI Plagiarism', *The Serials Librarian*, 85(5-6), pp. 181–189. Available at: https://doi.org/10.1080/0361526x.2024.2433256.

Guo, K., Wang, J. and Chu, S.K.W. (2022) 'Using chatbots to scaffold EFL students' argumentative writing', *Assessing Writing*, 54, 100666. Available at: https://doi.org/10.1016/j.asw.2022.100666.

Guo, K. and 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), pp. 8435–8463. Available at: https://doi.org/10.1007/s10639-023-12146-0.

Hadra, M., Cambridge, K. and Mesbah, M. (2026) 'Evaluating the accuracy and reliability of AI content detectors in academic contexts', *International Journal for Educational Integrity*, 22(1), 4. Available at: https://doi.org/10.1007/s40979-026-00213-1.

Han, J. and Li, M. (2024) 'Exploring ChatGPT-supported teacher feedback in the EFL context', *System*, 126, 103502. Available at: https://doi.org/10.1016/j.system.2024.103502.

Hattie, J. and Timperley, H. (2007) 'The Power of Feedback', *Review of Educational Research*, 77(1), pp. 81–112. Available at: https://doi.org/10.3102/003465430298487.

Hyland, K. and Hyland, F. (2006) 'Feedback on second language students' writing', *Language Teaching*, 39(2), pp. 83–101. Available at: https://doi.org/10.1017/s0261444806003399.

Hyland, K. and Hyland, F. (eds) (2019) *Feedback in second language writing: Contexts and issues*. 2nd edn. Cambridge: Cambridge University Press. Available at: https://doi.org/10.1017/9781108635547.

Hyland, K. (2025) 'ESP and writing', in Paltridge, B. and Starfield, S. (eds) *The handbook of English for specific purposes*. 2nd edn. Hoboken, NJ: Wiley-Blackwell, pp. 89–106. Available at: https://doi.org/10.1002/9781119985068.ch5.

Ibrahim, K. (2023) 'Using AI-based detectors to control AI-assisted plagiarism in ESL writing: "The Terminator Versus the Machines"', *Language Testing in Asia*, 13(1), 46. Available at: https://doi.org/10.1186/s40468-023-00260-2.

Imran, M. and Almusharraf, N. (2023) 'Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature', *Contemporary Educational Technology*, 15(4), ep464. Available at: https://doi.org/10.30935/cedtech/13605.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023) 'ChatGPT for good? On opportunities and challenges of large language models for education', *Learning and Individual Differences*, 103, 102274. Available at: https://doi.org/10.1016/j.lindif.2023.102274.

Kawashima, T. (2023) 'Student perceptions of Grammarly, teacher's indirect and direct feedback: Possibility of machine feedback', *The JALT CALL Journal*, 19(1), pp. 113–139. Available at: https://doi.org/10.29140/jaltcall.v19n1.1017.

Kim, J., Yu, S., Detrick, R. and Li, N. (2024) 'Exploring students' perspectives on Generative AI-assisted academic writing', *Education and Information Technologies*, 30(1), pp. 1265–1300. Available at: https://doi.org/10.1007/s10639-024-12878-7.

Kohnke, L., Moorhouse, B.L. and Zou, D. (2023) 'ChatGPT for Language Teaching and Learning', *RELC Journal*, 54(2), pp. 537–550. Available at: https://doi.org/10.1177/00336882231162868.

Koltovskaia, S. (2020) 'Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study', *Assessing Writing*, 44, 100450. Available at: https://doi.org/10.1016/j.asw.2020.100450.

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

Kosmyna, N., Hauptmann, E., Yuan, Y.T., Situ, J., Liao, X.-H., Beresnitzky, A.V., et al. (2025) 'Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task', *arXiv* [Preprint], arXiv:2506.08872. Available at: https://doi.org/10.48550/arXiv.2506.08872.

Lea, M.R. and Street, B.V. (1998) 'Student writing in higher education: An academic literacies approach', *Studies in Higher Education*, 23(2), pp. 157–172. Available at: https://doi.org/10.1080/03075079812331380364.

Lendvai, G.F. (2025) 'ChatGPT in Academic Writing: A Scientometric Analysis of Literature Published Between 2022 and 2023', *Journal of Empirical Research on Human Research Ethics*, 20(3), pp. 131–148. Available at: https://doi.org/10.1177/15562646251350203.

Li, J., Link, S. and Hegelheimer, V. (2015) 'Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction', *Journal of Second Language Writing*, 27, pp. 1–18. Available at: https://doi.org/10.1016/j.jslw.2014.10.004.

Li, J., Zong, H., Wu, E., Wu, R., Peng, Z., Zhao, J., et al. (2024) 'Exploring the potential of artificial intelligence to enhance the writing of english academic papers by non-native english-speaking medical students - the educational application of ChatGPT', *BMC Medical Education*, 24(1), 736. Available at: https://doi.org/10.1186/s12909-024-05738-y.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. and Zou, J. (2023) 'GPT detectors are biased against non-native English writers', *Patterns*, 4(7), 100779. Available at: https://doi.org/10.1016/j.patter.2023.100779.

Link, S., Mehrzad, M. and Rahimi, M. (2020) 'Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement', *Computer Assisted Language Learning*, 35(4), pp. 605–634. Available at: https://doi.org/10.1080/09588221.2020.1743323.

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

Marzuki, Widiati, U., Rusdin, D., Darwin and 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. Available at: https://doi.org/10.1080/2331186x.2023.2236469.

Meyer, J., Jansen, T., Schiller, R., Liebenow, L.W., Steinbach, M., Horbach, A. and Fleckenstein, J. (2024) 'Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions', *Computers and Education: Artificial Intelligence*, 6, 100199. Available at: https://doi.org/10.1016/j.caeai.2023.100199.

Mizumoto, A. and Eguchi, M. (2023) 'Exploring the potential of using an AI language model for automated essay scoring', *Research Methods in Applied Linguistics*, 2(2), 100050. Available at: https://doi.org/10.1016/j.rmal.2023.100050.

Nazari, N., Shabbir, M.S. and Setiawan, R. (2021) 'Application of Artificial Intelligence powered digital writing assistant in higher education: randomized controlled trial', *Heliyon*, 7(5), e07014. Available at: https://doi.org/10.1016/j.heliyon.2021.e07014.

Nguyen, A., Hong, Y., Dang, B. and Huang, X. (2024) 'Human-AI collaboration patterns in AI-assisted academic writing', *Studies in Higher Education*, 49(5), pp. 847–864. Available at: https://doi.org/10.1080/03075079.2024.2323593.

Ofem, U.J., Owan, V.J., Iyam, M.A., Udeh, M.I., Anake, P.M. and Ovat, S.V. (2024) 'Students' perceptions, attitudes and utilisation of ChatGPT for academic dishonesty: Multigroup analyses via PLS–SEM', *Education and Information Technologies*, 30(1), pp. 159–187. Available at: https://doi.org/10.1007/s10639-024-12850-5.

Ozfidan, B., El-Dakhs, D.A.S. and Alsalim, L.A. (2024) 'The use of AI tools in English academic writing by Saudi undergraduates', *Contemporary Educational Technology*, 16(4), ep527. Available at: https://doi.org/10.30935/cedtech/15013.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., et al. (2021) 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews', *BMJ*, 372, n71. Available at: https://doi.org/10.1136/bmj.n71.

Park, H., Jeong, G. and Kim, B. (2026) 'Style as a Confound: False Positives in AI Detection of Non-Native Academic Writing', *arXiv* [Preprint], arXiv:2608.26710. Available at: https://doi.org/10.48550/arXiv.2608.26710.

Perkins, M., Furze, L., Roe, J. and MacVaugh, J. (2024) 'The Artificial Intelligence Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment', *Journal of University Teaching and Learning Practice*, 21(06). Available at: https://doi.org/10.53761/q3azde36.

Priem, J., Piwowar, H. and Orr, R. (2022) 'OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts', *arXiv* [Preprint], arXiv:2205.01833. Available at: https://doi.org/10.48550/arXiv.2205.01833.

Qadir, S.M. and Omar, R.M. (2025) 'Digital literacy, technophobia, and technophilia: Exploring their roles in artificial intelligence integration into English language education', *E-Learning and Digital Media* [Advance online publication]. Available at: https://doi.org/10.1177/20427530251395634.

Rad, H.S., Alipour, R. and Jafarpour, A. (2023) 'Using artificial intelligence to foster students' writing feedback literacy, engagement, and outcome: a case of Wordtune application', *Interactive Learning Environments*, 32(9), pp. 5020–5040. Available at: https://doi.org/10.1080/10494820.2023.2208170.

Raheem, B.R. and Ghafar, Z.N. (2025) 'Understanding EFL learners' academic writing problems and the benefits of artificial intelligence tools: A descriptive review', *Journal of Language and Pragmatics Studies*, 4(3), pp. 187–193. Available at: https://doi.org/10.58881/jlps.v4i3.141.

Ranalli, J. (2018) 'Automated written corrective feedback: how well can students make use of it?', *Computer Assisted Language Learning*, 31(7), pp. 653–674. Available at: https://doi.org/10.1080/09588221.2018.1428994.

Ranalli, J. (2021) 'L2 student engagement with automated feedback on writing: Potential for learning and issues of trust', *Journal of Second Language Writing*, 52, 100816. Available at: https://doi.org/10.1016/j.jslw.2021.100816.

Risko, E.F. and Gilbert, S.J. (2016) 'Cognitive Offloading', *Trends in Cognitive Sciences*, 20(9), pp. 676–688. Available at: https://doi.org/10.1016/j.tics.2016.07.002.

Rozi, F., Larasaty, G., Wahyuni, S., Hapsari, C.T., Farida, L.A., Sutrisno, A., et al. (2026) 'Grammar Correction in AI-Assisted Business English Writing Instruction', *Vision: Journal for Language and Foreign Language Learning*, 14(1), pp. 85–102. Available at: https://doi.org/10.21580/vjv14i228794.

Selim, A.S.M. (2024) 'The Transformative Impact of AI-Powered Tools on Academic Writing: Perspectives of EFL University Students', *International Journal of English Linguistics*, 14(1), 14. Available at: https://doi.org/10.5539/ijel.v14n1p14.

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

Small, H. (1973) 'Co-citation in the scientific literature: A new measure of the relationship between two documents', *Journal of the American Society for Information Science*, 24(4), pp. 265–269. Available at: https://doi.org/10.1002/asi.4630240406.

Song, C. and 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. Available at: https://doi.org/10.3389/fpsyg.2023.1260843.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., et al. (2024) 'Comparing the quality of human and ChatGPT feedback of students' writing', *Learning and Instruction*, 91, 101894. Available at: https://doi.org/10.1016/j.learninstruc.2024.101894.

Stevenson, M. and Phakiti, A. (2014) 'The effects of computer-generated feedback on the quality of writing', *Assessing Writing*, 19, pp. 51–65. Available at: https://doi.org/10.1016/j.asw.2013.11.007.

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

Swales, J.M. (1990) *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Taheri, R., Nazemi, N., Pennington, S.E., Clark, J.A. and Dadgostari, F. (2025) 'Factors influencing educators' AI adoption: A grounded meta-analysis review', *Computers and Education: Artificial Intelligence*, 9, 100464. Available at: https://doi.org/10.1016/j.caeai.2025.100464.

Teng, M.F. (2024) '"ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing', *Computers and Education: Artificial Intelligence*, 7, 100270. Available at: https://doi.org/10.1016/j.caeai.2024.100270.

van Eck, N.J. and Waltman, L. (2009) 'Software survey: VOSviewer, a computer program for bibliometric mapping', *Scientometrics*, 84(2), pp. 523–538. Available at: https://doi.org/10.1007/s11192-009-0146-3.

Venkatesh, V., Morris, M.G., Davis, G.B. and Davis, F.D. (2003) 'User acceptance of information technology: Toward a unified view', *MIS Quarterly*, 27(3), pp. 425–478. Available at: https://doi.org/10.2307/30036540.

Verano-Tacoronte, D., Bolívar-Cruz, A. and Sosa-Cabrera, S. (2025) 'Are university teachers ready for generative artificial intelligence? Unpacking faculty anxiety in the ChatGPT era', *Education and Information Technologies*, 30(14), pp. 20495–20522. Available at: https://doi.org/10.1007/s10639-025-13585-7.

Vygotsky, L.S. (1978) *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Waltman, L., van Eck, N.J. and Noyons, E.C.M. (2010) 'A unified approach to mapping and clustering of bibliometric networks', *Journal of Informetrics*, 4(4), pp. 629–635. Available at: https://doi.org/10.1016/j.joi.2010.07.002.

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

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q. and 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. Available at: https://doi.org/10.1016/j.jslw.2023.101071.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., et al. (2023) 'Testing of detection tools for AI-generated text', *International Journal for Educational Integrity*, 19(1), 26. Available at: https://doi.org/10.1007/s40979-023-00146-z.

Wood, D., Bruner, J.S. and Ross, G. (1976) 'The role of tutoring in problem solving', *Journal of Child Psychology and Psychiatry*, 17(2), pp. 89–100. Available at: https://doi.org/10.1111/j.1469-7610.1976.tb00381.x.

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

Zawacki-Richter, O., Marín, V.I., Bond, M. and Gouverneur, F. (2019) 'Systematic review of research on artificial intelligence applications in higher education – where are the educators?', *International Journal of Educational Technology in Higher Education*, 16(1), 39. Available at: https://doi.org/10.1186/s41239-019-0171-0.

Zhang, Z. and Hyland, K. (2018) 'Student engagement with teacher and automated feedback on L2 writing', *Assessing Writing*, 36, pp. 90–102. Available at: https://doi.org/10.1016/j.asw.2018.02.004.

Zhang, Z. (2020) 'Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions', *Assessing Writing*, 43, 100439. Available at: https://doi.org/10.1016/j.asw.2019.100439.

Zhang, M. and Crosthwaite, P. (2025) 'More human than human? Differences in lexis and collocation within academic essays produced by ChatGPT-3.5 and human L2 writers', *International Review of Applied Linguistics in Language Teaching*, 64(2), pp. 733–760. Available at: https://doi.org/10.1515/iral-2024-0196.

Zupic, I. and Čater, T. (2014) 'Bibliometric Methods in Management and Organization', *Organizational Research Methods*, 18(3), pp. 429–472. Available at: https://doi.org/10.1177/1094428114562629.

Supplementary figures

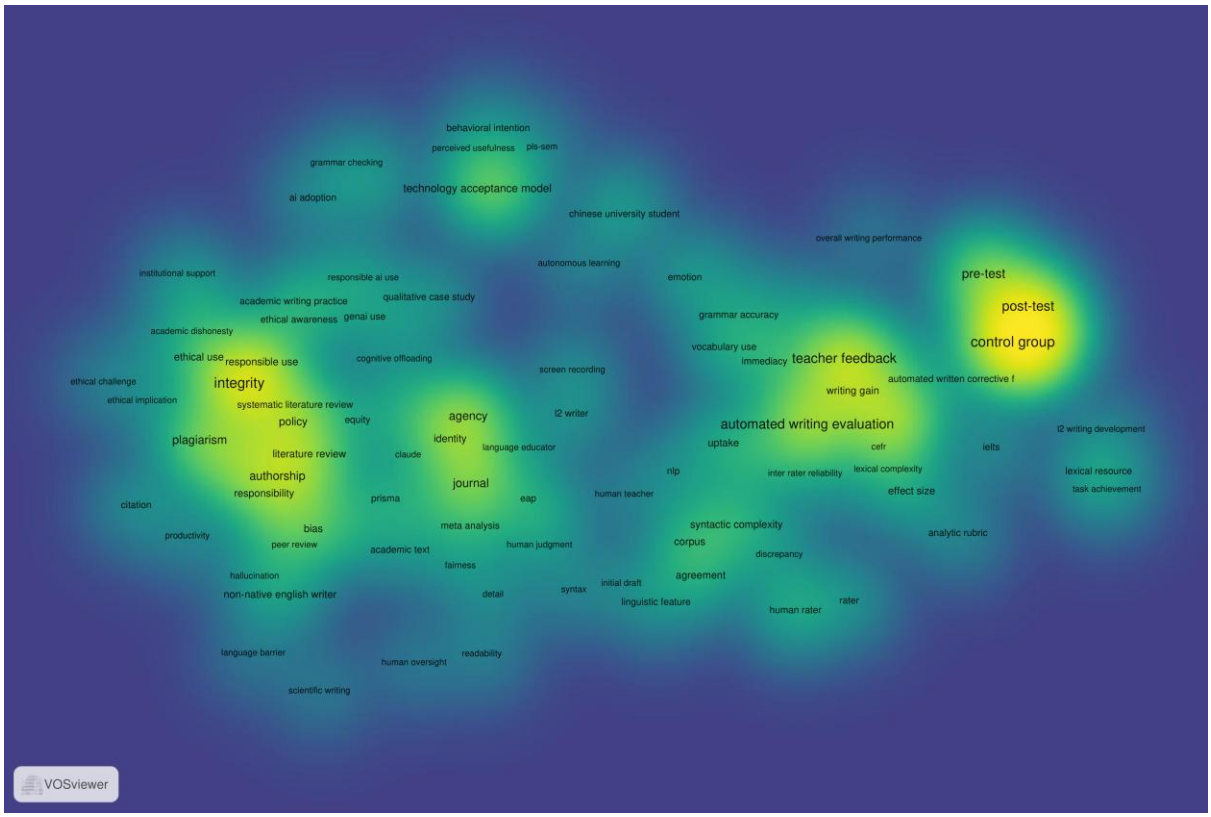


Figure S1. Density visualisation of the term co-occurrence map (VOSviewer).



Figure S2. Density visualisation of the reference co-citation map (VOSviewer).