Students' Reliance on AI Writing Tools in ESL/EFL Higher Education: A Systematic Review

Abdul Azim Mahda[1*], Nur Farah Fadhliah Mahmud[2], Sitti Hawa Sheikh Mohammad[3]

Academy of Language Studies, Universiti Teknologi MARA

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.102100214

Received: 10 October 2026; Accepted: 15 October 2026; Published: 24 October 2026

ABSTRACT

The systematic review presented herein investigates the reliance of students on artificial intelligence (AI) writing tools within higher education contexts for English as a second or foreign language (ESL/EFL). This review synthesises evidence on use, trust, overreliance, writing proficiency, and critical-thinking skills. Using PRISMA-informed procedures, a search of the Scopus database from January 2023 to 20 August 2026 yielded 611 unique English-language records. These records met the empirical eligibility criteria. The resulting investigation is reported as a systematic metadata-and-abstract review, as the export was limited to metadata and abstracts rather than a complete full-text corpus. Tactical reliance predominates among the synthesised data. AI is frequently used for grammar, vocabulary, paraphrasing, ideation, outlining, feedback, and revision, whereas whole-text generation functions primarily as a substitutive mechanism. Usefulness, ease of use, self-efficacy, and accessibility strongly predict initial adoption. Trust, conversely, is consistently qualified by documented concerns regarding inaccurate information, generic feedback, privacy risks, potential plagiarism, and the attenuation of authorial voice. Immediate gains in surface accuracy, fluency, organisation, revision efficiency, confidence, and writing self-efficacy represent the most consistent outcomes of AI support. Outcomes related to critical thinking remain highly conditional. Reasoning is supported through verification, reflective prompting, and guided comparison; uncritical, frequent use, by contrast, encourages cognitive offloading, diminishes originality, and weakens authorial control. The educational value of these tools ultimately depends on the specific tasks delegated to the system, the rigor of verification practices, and the pedagogical retention of learner agency.

Keywords: Artificial intelligence; AI writing tools; ESL; EFL; Reliance

INTRODUCTION

Generative AI is reshaping how university students learn to write in an additional language. Earlier automated writing evaluation (AWE) systems, grammar checkers, and paraphrasing applications mainly provided localised correction or reformulation. Since 2023, conversational systems such as ChatGPT have extended this support to ideation, outlining, drafting, explanation, feedback, and revision. For ESL/EFL students, these affordances can reduce linguistic difficulty and provide immediate assistance where teacher feedback is limited. At the same time, they can transfer planning, reasoning, and authorship decisions from the learner to the system.

Within this context, reliance cannot be understood as simple tool use. A learner who consults Grammarly for punctuation, evaluates a ChatGPT suggestion, and explains a revision engages with AI differently from a learner who submits a generated essay with little scrutiny. Both may report frequent use, but the educational consequences differ. In this review, reliance refers to the recurring delegation of writing-related tasks to an AI system, shaped by usefulness, trust, confidence, habit, task difficulty, access, and institutional expectations. Overreliance occurs when this delegation displaces learning-relevant cognitive work, verification, authorial judgment, or responsibility.

Existing evidence presents a similarly qualified picture. Experimental and mixed-methods studies commonly report improvements in writing scores, feedback quality, fluency, accuracy, organisation, confidence, or self-efficacy (Boudouaia et al., 2024; Guo et al., 2024; Mahapatra, 2024; Polakova & Ivenz, 2024). Other studies, however, report no significant advantage over teacher feedback, gains limited mainly to linguistic complexity, moderate-to-low feedback uptake, or no statistically significant improvement in broad critical-thinking measures (Alnemrat et al., 2025; Dong, 2024; Fan, 2023; Gao et al., 2025; Thi et al., 2023). Taken together, these findings indicate that access to AI does not determine learning alone. Task design, learner proficiency, feedback literacy, prompt quality, verification routines, and instructional guidance condition outcomes.

ESL/EFL higher education requires particular attention for three reasons. First, learners may benefit substantially from rapid lexical, grammatical, and stylistic support, making AI both attractive and potentially habit-forming. Second, linguistic insecurity may increase the apparent authority of fluent machine output even when its factual or rhetorical quality is weak. Third, academic writing in an additional language involves language development, disciplinary socialisation, critical reasoning, and identity formation. Consequently, an intervention that improves grammatical accuracy while weakening voice or evaluative judgment may enhance the product without developing the writer.

Against this background, the present review examines when reliance functions as productive scaffolding and when it becomes learning displacement. It maps the Scopus-indexed evidence, distinguishes bounded assistance from cognitive and product substitution, integrates writing-performance findings with evidence on trust and critical thinking, and develops a calibrated-reliance framework for research and pedagogy. Within this framework, task delegation, verification, and learner agency are central conditions for educationally productive AI use.

Research Questions

RQ1. How frequently, for what purposes, and at which stages of writing do ESL/EFL higher-education students use AI writing tools?

RQ2. What factors shape students' trust, continued use, dependency, and overreliance?

RQ3. What effects are associated with AI writing-tool use on writing performance, revision, confidence, autonomy, and related learning outcomes?

RQ4. Under what conditions does AI use support or undermine critical thinking, originality, metacognition, and authorial control?

RQ5. What pedagogical, methodological, and theoretical priorities follow from the evidence?

METHODOLOGY

Review Design and Reporting Approach

The present systematic review followed an explicit, reproducible protocol aligned with the reporting objectives of PRISMA 2020 (Page et al., 2021). As PRISMA functions fundamentally as a reporting guideline rather than a quality-appraisal instrument, its standard flow terminology required adaptation for a Scopus record set restricted to bibliographic metadata and abstracts. We then integrated descriptive evidence mapping with structured thematic synthesis. This methodological adjustment was necessitated by the substantial heterogeneity observed across the included studies regarding interventions, sample populations, specific technological instruments, measurement scales, and learning outcomes. Statistical pooling was therefore rendered inappropriate. Individual studies served as the primary unit of synthesis, ensuring that divergent findings were not artificially converted into a standardised effect size.

Information Source, Search Date, and Search Strategy

Scopus operated as the exclusive bibliographic database for this investigation. The systematic search was executed and exported on 20 August 2026, capturing literature published from January 2023 through the extraction date. This time frame was chosen to capture the widespread academic integration of public generative-AI writing systems in higher education. The advanced search query was structured around TITLE-ABS-KEY blocks representing four distinct conceptual domains: AI writing tools, ESL/EFL or second-language educational contexts, higher-education student populations, and outcomes related to reliance, usage patterns, writing proficiency, or critical thinking. The inclusion parameters deliberately encompassed specific named applications and broader technological categories, including ChatGPT, GPT models, large language models, Grammarly, QuillBot, generative AI (GenAI), automated writing evaluation (AWE) systems, automated feedback, and automated essay scoring (see Appendix A for the complete search string). The search was restricted to English-language materials published between 2023 and 2026 and indexed by Scopus as original articles, conference papers, or systematic reviews. The reviews were isolated during the initial identification phase for contextual orientation but strictly excluded from the primary empirical synthesis. No subject-area restrictions were imposed, as relevant empirical investigations span applied linguistics, education, educational technology, psychology, and composition studies.

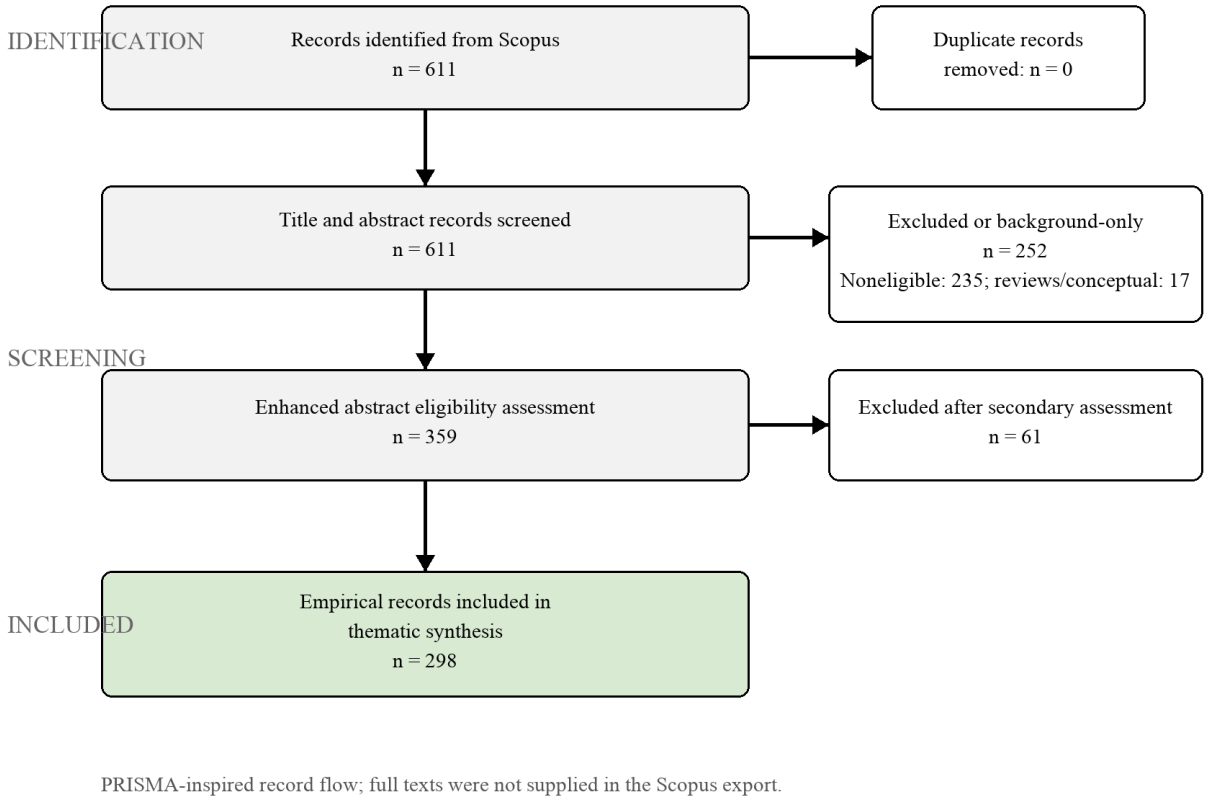
Eligibility Criteria

Stringent inclusion and exclusion parameters governed the title-and-abstract screening process. Studies were retained for synthesis only if they met five specific criteria. First, the participant sample was required to consist of university, college, tertiary, undergraduate, postgraduate, doctoral, or comparable higher-education learners. Second, the instructional environment had to involve ESL, EFL, L2 English writing, English for academic purposes, academic English, or English-medium higher education featuring an explicit second-language writing component. Third, the primary focal concept necessitated the direct, learner-facing application of an AI writing system for purposes such as ideation, drafting, linguistic support, paraphrasing, iterative feedback, assessment for learning, or textual revision. Fourth, eligible measured outcomes included usage frequency, adoption purpose, tool acceptance, user trust, technological dependency, systemic overreliance, writing performance metrics, critical thinking, cognitive offloading, learner autonomy, authorial agency, self-regulation, or metacognitive monitoring. Finally, the abstract was required to define an empirical research design or sample population explicitly. Conversely, records were systematically excluded if they examined primary or secondary school learners, focused exclusively on educators without yielding direct student evidence, or investigated non-writing applications such as speaking, reading, translation, materials development, AI detection algorithms, or model-scoring performance independent of learner interaction. Incidental references to AI systems within a broader context were deemed insufficient for inclusion. Studies in which an AWE system functioned solely to score an outcome measure, without direct learner engagement, were excluded for the same reason. Mixed-methods investigations involving both students and teachers were preserved exclusively when student practices or empirical outcomes were reported as distinct variables. Appendix B documents detailed eligibility thresholds.

Screening and Deduplication

The initial CSV export generated a corpus of 611 records containing 33 discrete bibliographic fields, including title, abstract, author keywords, indexed keywords, digital object identifier (DOI), document classification, publication year, source title, and cited-by metrics. Initial verification procedures utilising DOI and normalised-title cross-referencing identified zero duplicates within the dataset. The screening protocol subsequently advanced through two highly structured stages. The first stage evaluated all titles and abstracts against the established criteria for date range, publication language, target population, academic context, specific AI intervention, writing focus, target outcomes, and empirical methodology. This initial filtration excluded 235 records and segregated 17 review or conceptual manuscripts as supplementary background literature. In the second stage, 359 potentially eligible records underwent enhanced abstract eligibility verification. An additional 61 records were excluded because their abstracts failed to explicitly establish both direct student evidence and a focal ESL/EFL writing application. The final primary synthesis

therefore drew on 298 unique empirical records. To ensure analytical consistency across the extensive dataset, deterministic term matching was used during the initial screening phase. Borderline cases, incidental AI mentions, teacher-exclusive samples, and studies that used systems solely for automated scoring were then manually verified. Although explicit documentation of these protocols ensures reproducibility, a single reviewer conducted the entire screening procedure. The Limitations section formally acknowledges the potential for selection bias associated with this constraint.



PRISMA-inspired record flow; full texts were not supplied in the Scopus export.

Figure 1. PRISMA-Inspired Flow of Scopus Record Screening

Data Extraction and Coding

Data extraction procedures systematically isolated specific bibliographic and methodological variables for each included record, capturing the author, publication year, title, source, digital object identifier, document type, abstract, keywords, study design, geographical or educational context, and relevant outcome categories. Tool categories were coded non-exclusively to reflect the overlapping technological ecosystems learners inhabit. These categories included ChatGPT or specific Generative Pre-trained Transformer models; general generative artificial intelligence or large language models; automated writing evaluation systems providing corrective feedback; Grammarly; QuillBot; and a residual category for unspecified algorithmic writing tools. Study designs were similarly classified into mixed methods, experimental or quasi-experimental frameworks, survey or correlational models, qualitative investigations, longitudinal or action-research paradigms, and empirical studies lacking sufficient methodological detail within the abstract. Outcome coding utilised six non-exclusive categories spanning usage frequency and practices, acceptance and perceptions, trust and overreliance, writing performance, critical thinking and reasoning, and learner autonomy. This non-exclusive coding matrix was necessitated by the complex nature of the primary literature, where a single investigation frequently addressed both technological adoption and subsequent writing performance, or simultaneously evaluated critical thinking and textual output. The resulting quantitative counts, therefore, indicate the thematic visibility of each category rather than representing mutually exclusive groups of studies.

Quality and Strength-of-Evidence Appraisal

A design-sensitive appraisal framework replaced a standardised quality score to ensure interpretation consistency, which would otherwise obscure methodological variances across non-equivalent study designs. Evaluative confidence in individual claims was derived from specific indicators observable within the abstract parameters. These indicators included the precise articulation of the sample and methodology, the integration of multiple data sources such as student texts or behavioural logs, the presence of controlled comparators or repeated measures, the explicit reporting of statistical or thematic analyses, and the formal acknowledgement of study limitations. Greater interpretive weight was systematically assigned to convergent evidence drawn across diverse research designs rather than to isolated effect claims. Methodological caution dictated a strict conceptual separation between self-reported perceived improvements and objectively assessed writing performance. Cross-sectional technology acceptance models were strictly interpreted as correlational rather than indicative of causal mechanisms.

Synthesis

The subsequent synthesis progressed through the iterative comparison of coded evidence, adhering to the pattern-seeking principles of thematic analysis (Braun & Clarke, 2006). Deductive categories derived directly from the primary research questions provided the initial analytical scaffolding. Inductive comparison subsequently refined these boundaries, isolating distinct phenomena such as surface-language assistance, planning and ideation, feedback processing, whole-product substitution, trust calibration, cognitive offloading, integration of human feedback, and institutional constraints. Contradictory findings were deliberately retained and analysed rather than dismissed as statistical noise. While descriptive counts characterised the structural dimensions of the evidence base, substantive claims about educational mechanisms were drawn exclusively from convergence across process data, specific intervention outcomes, documented learner accounts, and comparative research designs.

RESULTS

Study Selection and Corpus Characteristics

The final corpus of 298 included records comprised 276 peer-reviewed journal articles and 22 conference papers. Publication volume exhibited a sharp upward trajectory reflecting the rapid disciplinary response to generative technologies. Specifically, nine qualifying studies were published in 2023, accelerating to 46 in 2024 and 120 in 2025, culminating in 123 studies published by the search cut-off date of 20 August 2026. This 2026 figure captures only a partial calendar year and should not be misconstrued as a finalised annual volume. Methodologically, mixed-methods studies dominated the corpus (n=109), followed sequentially by experimental or quasi-experimental frameworks (n=67), qualitative investigations (n=54), survey or correlational designs (n=38), longitudinal, case, or action-research models (n=20), and empirical studies presenting insufficient methodological specification in the abstract (n=10).

Technological tool categories demonstrated significant conceptual overlap across the literature. ChatGPT, or specifically named GPT models, appeared in 147 records, while general generative artificial intelligence or large-language-model terminology appeared in 126 instances. Traditional automated writing evaluation systems providing written corrective feedback accounted for 57 records, followed by specific commercial applications including Grammarly in 34 studies and QuillBot in 16 studies. In comparison, 21 records examined an alternate or incompletely specified algorithmic writing tool. The thematic outcome map similarly reflected complex overlaps in the research. The literature prioritised writing performance or specific skills (209 records), closely followed by technology acceptance or learner perceptions (199 records), and usage frequency, purpose, or practices (158 records). Examinations of user trust, dependency, or systemic overreliance appeared in 95 records, while issues of autonomy or self-regulation appeared in 89. Critical thinking or analytical reasoning constituted the smallest focal category, appearing in exactly 43 records. These mapped distributions indicate that the current empirical literature remains heavily weighted toward documenting perceived utility and immediate textual writing outcomes. Conversely, the structural

impacts of technological dependency and the mediation of critical thinking are emerging as secondary but rapidly developing investigative streams.

Table 1. Characteristics of the Included Evidence Base

Dimension	Category	n
Records	Included empirical records	298
Publication year	2023 / 2024 / 2025 / 2026 through 20 August	9 / 46 / 120 / 123
Document type	Journal articles/conference papers	276 / 22
Design	Mixed methods	109
Design	Experimental or quasi-experimental	67
Design	Qualitative	54
Design	Survey or correlational	38
Design	Longitudinal, case, or action research	20
Design	Empirical, design detail incomplete in abstract	10
Outcome	Writing performance or skills	209
Outcome	Acceptance or perceptions	199
Outcome	Frequency, purpose, or practices	158
Outcome	Trust, dependency, or overreliance	95
Outcome	Autonomy or self-regulation	89
Outcome	Critical thinking or reasoning	43

Note. Tool and outcome codes were non-exclusive; counts can exceed 298.

Evidence Matrix Overview

An examination of the constructed evidence matrix reveals a discernible thematic trajectory within the recent literature. Initial scholarly attention heavily prioritised technology acceptance models. Current empirical investigations demonstrate a marked transition toward complex cognitive and authorial mechanisms, specifically targeting textual revision processes, the calibration of user trust, cognitive offloading, the preservation of authorial voice, output verification, and pedagogically guided integration. Table 2 encapsulates this methodological and thematic variation. This table serves as an illustrative cross-section of the corpus rather than a secondary analytical filter.

Table 2. Representative Evidence Across the Reliance Continuum

Study	Design	Main evidence	Qualification
Pinninti (2024)	Survey	Regular generation/review; cautious trust	Single class; self-report
Thi (2024)	Mixed data	Language refinement, homework, self-study; overreliance concerns	n = 53 survey; five interviews
Dong (2024)	Mixed methods	Acceptance predicted behaviour; use predicted complexity only	Teacher support requested
Mahapatra (2024)	Intervention	Significant academic-writing improvement	Short-term tertiary ESL setting
Fan (2023)	Quasi-experimental mixed methods	No significant Grammarly advantage over teacher feedback	Two intact classes
Zou et al. (2026)	Comparative	Teacher feedback: greater uptake; ChatGPT: organisational strength	Small Chinese sample
Hong and Shin (2025)	Three-condition comparison	Guided and free use exceeded restriction; guided use strongest for content	Same instructor; n = 103
Ruyang et al.	Mixed-methods	Verification norms supported	Two classrooms

(2026)	cases	deeper revision and transfer	
Tekir (2026)	Mixed methods	Frequent reliance linked to fluency but lower originality/reasoning	Optional access; n = 78
Hu et al. (2025)	Mixed methods	Perceived CT gains; limited higher-order and meta-CT gains	Three EAP writing tasks
Gao et al. (2025)	Quasi-experimental	No significant overall CT or self-efficacy difference	Seven-week intervention
Alshehri et al. (2026)	Mixed-methods quasi-experiment	Guided verification improved writing and digital CT	Intact classes; contextual limits

FINDINGS

RQ1. Frequency and Purposes of AI Writing Tool Application

Empirical data demonstrate that learners distribute AI-assisted writing tasks along a functional continuum rather than relying on automated systems for a single, uniform purpose. This distribution ranges from localised linguistic correction to systemic content transformation and complete product generation. Lower-order applications encompass grammar verification, vocabulary refinement, orthographic correction, paraphrasing, and stylistic adjustment. Conversely, higher-order delegations include conceptual brainstorming, knowledge inquiry, structural outlining, argument synthesis, and iterative revision. English majors in Vietnam, for example, used AI mainly for independent study and targeted language refinement rather than relying on algorithmic assistance for conceptual organisation (Thi, 2024). Similarly, Indonesian master's students reported a combination of vocabulary enhancement, grammar correction, and structural generation, which they actively supplemented through peer consultation and primary-source verification (Werdiningsih et al., 2024b). Thai English majors identified paraphrasing, ideation, and coherence enhancement as their primary operational objectives (Tanharat & Phootirat, 2025).

Usage frequency exhibits substantial contextual variation. A large Turkish undergraduate sample reported infrequent ChatGPT use during L2 writing despite high baseline perceptions of utility; statistical analysis confirmed a positive correlation between these perceptions and actual use (Aksakallı & Daşer, 2025). An Indian ESL cohort presented a contrasting pattern. These learners engaged in regular text generation and review, documenting explicit intentions to increase future use (Pinninti, 2024). This divergence invalidates assumptions of uniform technological dependency among contemporary student populations. Algorithmic application is strictly mediated by institutional access, course parameters, technological familiarity, disciplinary norms, baseline proficiency, and the formal legitimacy granted by local academic policies.

Process tracing yields greater analytical clarity than basic frequency metrics. The introduction of pre-writing support systematically expanded documented planning strategies (Nguyen et al., 2025). Iterative revision studies also found significantly higher uptake of localised linguistic feedback than of structural or conceptual recommendations (Cengiz et al., 2025). Isolated prompt counts cannot distinguish sustained pedagogical engagement from repeated cognitive outsourcing. They cannot identify which specific stages of writing remain under active learner control.

Generative algorithms function as direct substitutes for learner output. Screen-recorded independent sessions with Saudi students revealed a bifurcation between a process-oriented methodology spanning multiple writing stages and a product-oriented strategy that relies entirely on machine-generated essays processed through an algorithmic humaniser (Dahmash, 2026). This product-oriented pattern fundamentally diverges from targeted correctional consultation. It actively obscures the visibility of student reasoning and invalidates the final textual product as an accurate metric of independent linguistic capability. Assessing reliance requires evaluating the specific cognitive labour delegated, not simply measuring tool selection or interaction duration.

RQ2. Determinants of Trust, Dependency, and Overreliance

Technology acceptance models exhibit consistent predictive patterns across the literature. Perceived usefulness and ease of use reliably predict learner attitudes, behavioural intentions, and sustained engagement, often alongside task relevance, domain self-efficacy, and facilitating institutional conditions. An extended technology acceptance model explained 71% of the variance in Chinese EFL learners' explicit intentions to use Grammarly, with usefulness, ease of application, enjoyment, and task relevance as the primary predictors (Liang et al., 2024). Usefulness, ease, and user satisfaction demonstrated strong associations with continuance intention among Kazakhstani TEFL students. Qualitative open responses from this cohort simultaneously documented severe anxieties regarding factual reliability, academic plagiarism, and the progressive degradation of independent writing proficiency (Bekturova et al., 2025). Integrated theoretical frameworks, including task-technology fit and UTAUT models, identified mathematically comparable predictive structures within GenAI-assisted L2 writing environments (Hu & Gong, 2025; Yang, 2025).

High perceived utility does not neutralise risk awareness. A comprehensive Chinese survey identified a distinct usefulness-risk paradox: students reporting high perceived utility also showed acute awareness of algorithmic risks. Positive operational attitudes predicted both beneficial and detrimental perceived outcomes (Li et al., 2026). This cognitive duality explains why frequent users remain highly cautious evaluators. Algorithmic trust operates as a dynamically calibrated metric rather than a static psychological disposition. Learners readily trust automated systems for localised grammar correction and brainstorming but actively distrust algorithmic outputs on disciplinary epistemologies, formal citations, cultural nuance, and evaluative feedback.

Qualitative data isolate the precise mechanisms of trust breakdown. Students systematically identify fabricated information, improper citation formatting, unnecessary linguistic complexity, generic rhetorical advice, data privacy vulnerabilities, unequal premium access, plagiarism hazards, and the systemic erasure of personal authorial voice (Kohnke, 2024; Tanharat & Phootirat, 2025; Xu & Jumaat, 2024). Elaborate conversational explanations can paradoxically cause GenAI to appear more authoritative than traditional diagnostic checkers (Kohnke, 2024). Highly polished linguistic output frequently functions as a deceptive signal of epistemic reliability for language learners. Pedagogical interventions must explicitly untangle linguistic fluency from factual dependability.

Direct empirical quantifications of systemic overreliance remain scarce. A survey of Saudi Grammarly users classified 37.7% of participants as exhibiting high overreliance, leaving 62.3% in the low or no overreliance categories. Statistical modelling revealed no significant association between perceived usefulness and actual overreliance (Abdullah et al., 2026). This finding contradicts the assumption that highly functional tools automatically induce learner dependency. The synthesised data indicate that task architecture, established verification routines, acute assessment pressure, self-regulatory capacity, and the immediate availability of human instructional support strictly condition overreliance.

Established classroom norms directly modulate trust calibration. A comparative case study demonstrated that explicit performance expectations, learning-oriented accountability structures, and the deliberate positioning of AI as contingent rather than authoritative input systematically produced deeper textual revision, rigorous output verification, and stronger skill transfer. Environments characterised by extreme assessment pressure and deficient verification protocols drove students to accept algorithmic feedback uncritically, resulting in documented patterns of learning displacement (Ruyang et al., 2026). Calibrated trust functions as a measurable pedagogical outcome rather than an innate learner characteristic.

RQ3. Effects Associated with AI-Supported Writing

The corpus synthesis shows concentrated short-term performance gains when AI serves as formative feedback or a guided revision tool. Documented improvements manifest across grammatical accuracy, syntactic conciseness, structural organisation, key information integration, revision efficacy, task confidence, and the reduction of writing anxiety. Mixed-methods research confirmed significant statistical improvements

in academic writing performance following iterative ChatGPT feedback (Mahapatra, 2024). A quasi-experimental design incorporating 110 participants similarly documented measurable gains in conciseness, grammatical precision, information retention, and the strategic deployment of passive voice (Polakova & Ivenz, 2024). The integration of AI within a peer-feedback framework systematically elevated both the diagnostic quality of student-generated feedback and subsequent writing proficiency (Guo et al., 2024). Targeted QuillBot interventions produced corresponding enhancements in textual performance, reduced linguistic apprehension, and elevated writing self-efficacy (El-Garawany, 2024).

These performance gains remain highly concentrated within lower and middle-level writing dimensions. Behavioural analysis within a mixed-methods framework showed that ChatGPT interaction reliably predicted specific vectors of linguistic complexity but did not influence the broader spectrum of accuracy, fluency, and holistic complexity outcomes (Dong, 2024). A controlled Grammarly intervention yielded no statistically significant advantage over traditional teacher feedback regarding syntactic complexity, lexical variation, accuracy, or fluency (Fan, 2023). A comparative evaluation of Jordanian learners indicated that while both algorithmic and human feedback improved argumentative writing, the performance differential between the two feedback sources remained statistically insignificant and the aggregate effect size was small (Alnemrat et al., 2025). These outcomes invalidate assumptions of categorical algorithmic superiority over human instructional feedback.

Feedback uptake metrics strongly indicate functional complementarity. Chinese undergraduates demonstrated higher engagement and executed more accurate revisions when using teacher feedback, although ChatGPT proved structurally advantageous for macroscopic text organisation (Zou et al., 2026). A parallel comparison found higher revised-essay scores after ChatGPT feedback, even though learners rated teacher feedback as more comprehensible and practically useful. These students prioritised human clarity and emotional scaffolding, while describing ChatGPT diagnostics as exhaustive yet often generic and overly complex (Han & Wang, 2026). This divergence requires the analytical separation of objective performance metrics from subjective learner experiences. A specific feedback source may drive textual revision without becoming the preferred or most cognitively accessible instructional relationship.

The intensity of instructional guidance directly dictates performance outcomes. Korean college students operating under guided ChatGPT access statistically outperformed cohorts restricted from using the tool, with the guided group achieving the highest metrics in overall content generation (Hong & Shin, 2025). The structural addition of reflective mechanisms to AI-supported writing protocols significantly improved textual performance, self-efficacy, self-regulation, and the management of cognitive load when compared against unstructured AI access (Liu et al., 2023). Automated writing evaluation systems maximise their educational value only when learners receive explicit operational training, conduct mandatory suggestion comparisons, retain ultimate revision responsibility, and actively hybridise automated diagnostics with human input (Dizon & Gold, 2023; Fan, 2023; Mekheimer, 2025).

Algorithmic tools measurably improve textual outputs, but these improvements are neither guaranteed nor uniformly distributed across the cognitive domains of writing. Surface accuracy and revision speed improve consistently. Disciplinary argumentation, evidentiary quality, rhetorical judgment, and transferable independent skills do not. Short-term interventional designs currently fail to determine whether observed linguistic gains persist following the withdrawal of the algorithmic scaffolding.

RQ4. Conditions Governing Critical Thinking and AI Use

The interaction between algorithmic assistance and critical thinking perfectly illustrates the conditionality of educational effects. Structured algorithmic delegation reduces lower-level linguistic load, freeing cognitive capacity for advanced analysis, evaluation, and conceptual reflection. Students participating in an English for Academic Purposes study successfully deployed AI to locate literature, stimulate divergent reasoning, expose logical inconsistencies, synthesise discrete information, and prompt metacognitive reflection. These learners reported substantial gains in analytical reasoning, evaluative logic, and contextual relevance, though improvements in conceptual profundity, cognitive flexibility, and meta-critical thinking remained

statistically inconsistent (Hu et al., 2025). A semester-long collaborative inquiry demonstrated that the strategic coordination of human and generative resources across multiple argumentative stages yielded measurable improvements in both structural discourse and final written products (Chen et al., 2025).

Interventions explicitly oriented toward reasoning mechanisms consistently outperformed unrestricted tool access. Reflective scaffolding elevated both textual quality and self-regulatory behaviours (Liu et al., 2023). The systematic integration of critical-thinking prompts, hybridised AWE-GenAI feedback models, and mandatory verification workflows generated concrete gains in analytical evaluation and digital critical thinking (Alshehri et al., 2026; Hao et al., 2026; Yin & Dou, 2025). Algorithmic tools generate maximal educational value when task architectures force learners to interrogate machine outputs rather than passively receive them.

Contradictory evidence isolates the specific dangers of cognitive offloading. A seven-week chatbot intervention showed no statistically significant shifts in aggregate critical thinking or writing self-efficacy, limiting observable improvements to basic language construction (Gao et al., 2025). During timed EFL writing assessments, high-frequency AI reliance correlated strongly with textual fluency and structural organisation but also predicted reduced conceptual originality, weaker argumentation, severe cognitive offloading, and the systemic collapse of authorial control. Highly selective users successfully retained structural reasoning, deliberate planning, reflective capacity, and distinct authorial voice (Tekir, 2026). Cognitive offloading is neither inherently advantageous nor categorically detrimental. Delegating orthographic correction or basic conceptual enumeration successfully releases working memory for higher-order evaluation. Conversely, delegating central claims, logical reasoning, and final rhetorical phrasing actively bypasses the very cognitive mechanisms through which critical thinking develops.

Explicit cognitive-offload instruction resolves this pedagogical tension. Students who deliberately delegated lower-order mechanical tasks while remaining strictly accountable for higher-order critique, systemic revision, peer collaboration, and metacognitive reflection improved both essay quality and formal critical-thinking performance (Hong et al., 2025). The causal mechanism driving this improvement was the structured pedagogical redistribution of cognitive labour, not the mere presence of the algorithm. Analysis must abandon binary inquiries about whether AI diminishes thinking and focus instead on which specific operations are delegated, which are retained, and what systematic verification protocols follow the generation event.

RQ5. Pedagogical Priorities Emerging from Empirical Evidence

The preservation of the student as the principal decision-making agent emerges as the primary pedagogical priority across all methodological paradigms. Process-oriented learners deploy AI strictly as a contingent resource, cross-referencing generated suggestions against primary goals and independently rejecting incongruous outputs. Product-oriented users passively absorb generated language and complete text blocks absent critical evaluation. The former operational pattern supplies cognitive scaffolding. The latter enacts complete learning substitution by masking linguistic deficits behind machine-generated fluency. Authorial voice is uniquely vulnerable within L2 writing contexts because algorithmic suggestions frequently exceed the student's current lexical repertoire. While learners value this idiomatic support, they consistently report that the resulting outputs introduce inappropriate cultural mismatches, excessive syntactic complexity, and a fundamental erasure of their personal academic identity (Cengiz et al., 2025; Werdiningsih et al., 2024a). Authentic authorial voice is preserved through selective uptake, the mandatory justification of textual alterations, and structured draft comparisons rather than the mere prohibition of copied syntax.

The retention and structural integration of human feedback represent a necessary corollary to algorithmic support. Students explicitly ask instructors for guidance on acceptable use parameters, how the evaluate complex algorithmic feedback, and specific disciplinary conventions (Dong, 2024; Wang, 2024). Learners structurally value teacher feedback for its emotional scaffolding, contextual clarity, and strict attention to semantic meaning. AI systems provide contrasting benefits characterised by extreme processing speed, perpetual availability, diagnostic comprehensiveness, and the facilitation of low-stakes iterative loops (Han

& Wang, 2026; Zou et al., 2026). Peer collaboration successfully stimulates open-ended conceptual reasoning that generative algorithms consistently fail to elicit autonomously (Hu et al., 2025). This evidence mandates the design of a hybrid feedback ecology rather than attempting the wholesale algorithmic replacement of human instructional networks.

Institutional conditions heavily dictate the realities of technological reliance and structural equity. Access to premium algorithmic models, device availability, disciplinary conventions, implicit teacher messaging, and assessment architectures directly control both tool utilisation and learner disclosure. Documented anxieties about data privacy and unequal technological access now parallel standard concerns about academic plagiarism and factual hallucination (Kohnke, 2024; Xu & Jumaat, 2024). Unmitigated digital inequality rapidly translates into structural writing inequality when affluent demographic segments secure exclusive access to highly capable generative models and superior diagnostic feedback. Disclosure norms remain critically unsettled within academic environments. While L2 doctoral students articulate strong theoretical support for methodological transparency in academic publishing, nearly 60% admit to actively concealing their use of GenAI (Wu et al., 2026). This behavioural gap reflects acute reputational risk calculations and highly inconsistent institutional policies, not innate academic dishonesty. Assessment architectures that universally prohibit algorithmic assistance while everyday academic study normalises it actively incentivise concealment. The empirical evidence necessitates assessment frameworks designed to render the writing process visible through mandatory prompt logs, annotated revision tracking, oral defence of generated text, source verification requirements, and the explicit comparison of independent versus assisted performance. These frameworks recognise that surveillance technologies cannot mathematically prove learning. Institutions must instead require students to document their authorial decisions and set precise parameters for permissible assistance.

Tool affordances and baseline learner proficiency function as the primary moderators of technological reliance. Corrective systems support localised paraphrasing, whereas advanced generative models intervene across the entire planning, drafting, and revision continuum. Lower-proficiency learners gain substantial immediate benefits from heavy linguistic support but often lack the domain knowledge to identify logically weak or factually hallucinated suggestions. Higher-proficiency cohorts evaluate automated feedback with greater strategic precision, although their actual uptake remains highly selective (Alnemrat et al., 2025; Thi et al., 2023). Variations in study design actively drive the inconsistent conclusions present in the literature. Perception surveys accurately map perceived utility but fail to measure objective improvement. Short-term interventions capture immediate performance spikes but rarely demonstrate longitudinal skill transfer. Qualitative process-tracing isolates the critical threshold where students either critically evaluate or blindly adopt algorithmic output. While the literature has shifted from basic acceptance modelling toward investigating trust calibration, cognitive offloading, and authorial voice, subsequent research demands stable theoretical constructs and replicable task architectures that can withstand rapid model updates.

DISCUSSION

A Calibrated-Reliance Framework

The systematic synthesis of 298 Scopus-indexed empirical records on ESL/EFL higher education reveals a fundamental analytical conclusion: algorithmic reliance resists adequate representation through single-metric frequency measures. Educational consequences manifest precisely through the interaction of task delegation, trust calibration, output verification, and the preservation of human agency. The proposed calibrated-reliance framework models this relationship as a sequence of steps. A specific learner's need initiates task delegation. This delegation is subsequently filtered through the learner's perception of utility and calibrated trust. The learner then accepts, modifies, rejects, or formally verifies the algorithmic output. This sequence of actions determines whether the interaction supports skill development or displaces necessary cognitive labour.

Data addressing RQ1 and RQ2 demonstrate that while students apply algorithmic tools across the entire writing continuum, the most common reliance vectors remain bounded: localised language support,

structural paraphrasing, initial ideation, diagnostic feedback, and textual revision. Initial adoption correlates strongly with perceived usefulness and ease of application. However, user trust functions strictly as a task-specific variable heavily qualified by acute concerns regarding factual reliability, conceptual originality, data privacy, and the preservation of authorial voice. Four primary modes of reliance organise these observed patterns, which are within assistive reliance, learners delegate bounded, surface-level operations, notably orthographic verification or grammatical diagnosis. Users deploy algorithms for conceptual brainstorming, alternative explanations, structural outlining, or diagnostic feedback during generative scaffolding while strictly retaining academic responsibility for evidentiary support and final expression. Collaborative regulation occurs when learners deliberately synthesise automated feedback with teacher and peer input, followed by mandatory verification and metacognitive reflection. Conversely, substitutive reliance manifests when students delegate core logical reasoning, primary drafting, or fundamental authorship, ultimately accepting the generated output absent critical scrutiny. These modes do not constitute fixed psychological learner typologies. A single student frequently transitions among these modes depending on immediate deadlines, task difficulty, institutional policy, domain confidence, or localised assessment pressure.

The framework successfully reconciles several contradictory findings associated with RQ3 and RQ4. High-frequency application frequently correlates with positive technological perceptions and perceived skill improvement, yet intensive algorithmic reliance simultaneously predicts diminished conceptual originality and severe cognitive offloading. This duality exists because raw frequency metrics fail to specify the nature of the delegated task or the depth of subsequent cognitive engagement. Algorithmic feedback routinely generates higher mathematical revision scores, even as students explicitly prioritise human teacher feedback because objective performance, instructional clarity, affective recognition, and interpersonal trust operate as distinct variables. Guided and unrestricted access models generally outperform total prohibition paradigms, but explicit instructional guidance yields superior content outcomes. Therefore, the critical analytical comparison is not a binary evaluation of AI presence versus absence. Research must instead contrast structured versus unstructured cognitive delegation and systematically verified versus unverified uptake of output.

As for RQ5, the synthesised evidence confirms that pedagogical value is not an automatic byproduct of technological access. Value generation depends entirely upon instructional frameworks that explicitly delineate task boundaries, mandate verification protocols, require rigorous feedback evaluation, and enforce authorial responsibility. This conditional perspective supports pedagogies that emphasise guided integration, process-visible assessment architectures, hybrid feedback ecologies, and proficiency-sensitive algorithmic support. This perspective concurrently explains why the educational function of generative AI must remain conceptualised as a mediated writing resource rather than an autonomous substitute for independent learner judgment.

Theoretical Implications

Traditional technology acceptance models retain descriptive utility but lack pedagogical sufficiency. Theoretical frameworks such as TAM and UTAUT predict operational intention via perceived usefulness, ease of use, social influence, facilitating conditions, domain self-efficacy, and user satisfaction. However, they fail to establish whether the resulting usage preserves or degrades actual learning. Comprehensive models of academic reliance must incorporate the structural variables of task criticality, formal verification, feedback literacy, authorial agency, and assessment accountability. Sociocultural frameworks, models of self-regulated learning, and theories of cognitive offloading provide the scaffolding needed to explain the critical role of human mediation. These perspectives allow researchers to definitively distinguish the productive reduction of linguistic burden from the detrimental displacement of core reasoning practices. A robust theoretical model of AI-supported L2 writing requires the direct integration of technology acceptance, self-regulatory capacity, advanced feedback literacy, and strategic cognitive allocation.

Methodological Implications

Future empirical designs must strictly isolate final product quality from durable, unaided learning and

distinct relational outcomes, notably user trust and authorial agency. Methodologies require comprehensive reporting parameters, including the specific generative model, premium access tiers, task stages, intervention durations, precise prompt architectures, preliminary training, interaction frequencies, copying permissions, and established verification requirements. Objective behavioural logs, rigorous draft comparisons, and independent post-tests remain necessary to validate subjective self-reports and separate productive assistance from wholesale cognitive substitution. Given the extreme velocity of technological evolution, the discipline would benefit from preregistered systematic updates and the maintenance of public evidence matrices to ensure longitudinal continuity.

Practical Implications for Teaching and Assessment

Instructional paradigms must prioritise task-level strategic guidance over generic pronouncements of encouragement or prohibition. Educators can deploy a delegation hierarchy effectively. Low-risk support parameters encompass basic error identification and localised alternative phrasing. Medium-risk operations include conceptual ideation, structural outlining, and diagnostic feedback that require mandatory cross-comparison. High-risk delegation involves the generation of core claims, evidentiary synthesis, data interpretation, logical argumentation, and complete textual drafting. These calculated risk levels require calibration against specific learning objectives. The pedagogical rules governing a basic grammar diagnostic should not replicate those governing a capstone research argument.

Verification mechanisms must become explicitly observable. Students can be required to tag algorithmic contributions formally, cross-reference generated outputs against authoritative primary sources, document actively rejected machine suggestions, justify substantial structural revisions, and retain comprehensive prompt or version histories. Instructors can mandate the inclusion of a structured AI-use statement detailing the specific tool, delegated task, applied verification methodology, and retained human intellectual contribution. This protocol transforms the act of disclosure from a punitive compliance exercise into a generative artifact of metacognitive evidence.

Instruction must systematically cultivate advanced feedback literacy. Students require explicit training to accurately identify generic, overly complex, factually inaccurate, or rhetorically misaligned machine suggestions. They must learn to translate automated diagnostics into their own linguistic repertoire, determine whether the feedback aligns with the assignment parameters, and empirically test whether a suggested revision genuinely enhances semantic meaning. Hybrid feedback sequences demonstrate substantial pedagogical promise. Algorithms supply rapid initial diagnostics, peer networks contribute authentic audience response and collaborative reasoning, and human instructors provide authoritative disciplinary judgment alongside selective, higher-order conceptual feedback.

Critical-thinking objectives must be structurally embedded directly within assignment prompts and task architectures. Students should face mandatory requirements to generate localised counterarguments, isolate absent evidentiary support, critically compare divergent algorithmic responses, identify logical inconsistencies, and explicitly justify the rejection of machine text. Instructors who merely instruct a system to "improve a text" actively invite passive cognitive acceptance. The assignment of structured collaborative roles such as primary proposer, designated sceptic, factual verifier, and final author effectively distributes cognitive activity while ensuring core responsibility remains with the human learner.

Institutional governance structures must actively address digital equity. If academic programs expect algorithmic integration, institutions must guarantee universal access to approved tools, provide explicit privacy guidance, and ensure the availability of meaningful non-digital alternatives. Academic integrity policies must conceptually separate proofreading, diagnostic feedback, linguistic translation, conceptual ideation, and whole-text generation, abandoning the treatment of all interactions as undifferentiated AI usage. Assessment architectures must strategically measure independent, unaided ability at controlled intervals while simultaneously providing explicit instruction in responsible, assisted practice.

Research Gaps and Future Agenda

Measurement of Reliance and Overreliance

The discipline currently lacks validated psychometric instruments that can distinguish high-frequency application from systemic cognitive dependence. Future measurement tools must assess the specific type of delegated task, unaided baseline performance, applied verification strategies, confidence calibration, the uncritical acceptance of inaccurate suggestions, and the quantified degradation of authorial voice. Subjective self-reports require mandatory triangulation against behavioural logs, sequential drafts, continuous screen recordings, concurrent think-aloud protocols, and independent post-tests. Overreliance must be demonstrated as an observable behavioural deficit rather than inferred mathematically from positive technological attitudes.

Longitudinal Transfer and Skill Retention

Current interventions rely predominantly on abbreviated timeframes. Immediate performance spikes in AI-assisted essays fail to establish durable skill transfer. Experimental designs require structured withdrawal phases, delayed post-tests, and alternating sequences of assisted and unassisted tasks to definitively determine whether learners internalise feedback principles and sustain cognitive gains across diverse genres and academic disciplines.

Dose, Task, and Proficiency Effects

Empirical research must isolate precise dose-response relationships and identify critical threshold effects across specific task stages, baseline proficiency levels, degrees of instructional guidance, and mandated verification requirements. Because a single generative tool can support structural planning while degrading source synthesis, studies must rigorously document tool versions, backend settings, exact prompts, temperature settings, and access tiers to enable accurate replication.

Objective Critical-Thinking and Authorship Outcomes

Research targeting critical thinking remains disproportionately limited compared to investigations of user perception and immediate writing performance. Future paradigms must analytically separate argument quality, evidentiary evaluation, conceptual originality, metacognitive monitoring, authorial voice, intellectual ownership, rhetorical decision-making, and longitudinal transfer. These variables demand independent, blinded evaluation alongside metrics of perceived subjective improvement.

Comparative Feedback Ecologies

Experimental comparisons of exclusive AI, exclusive teacher, exclusive peer, and structured hybrid feedback sequences require rigorous controls holding both task type and duration constant. These studies must systematically examine diagnostic accuracy, learner uptake rates, ultimate revision quality, sustained motivation, emotional scaffolding, institutional cost, and human teacher workload. The fundamental pedagogical inquiry remains identifying which specific feedback source should address each discrete level of writing and determining the optimal instructional sequence.

Context, Equity, and Policy

The current empirical corpus exhibits severe geographic concentration, and numerous abstracts fail to specify the educational context. Research must urgently expand into underrepresented geographic regions, complex multilingual institutions, resource-constrained academic settings, and disciplines beyond traditional English majors. These investigations must prioritise the structural impacts of premium-access inequality, technological accessibility, systemic data privacy risks, contradictory institutional governance, and the specific pedagogical effects of mandatory disclosure and process-based assessment architectures.

Table 3. Priority Research Agenda

Priority	Construct	Recommended design	Contribution
Dependence	Behavioural reliance scale	Logs, assisted/unassisted tasks, error acceptance	Distinguish use from inability to perform independently
Transfer	Durable learning	Delayed post-tests and AI-withdrawal designs	Separate improved product from learned skill
Critical thinking	Reasoning, originality, metacognition	Blinded rubric scoring plus process evidence	Identify productive versus harmful offloading
Feedback ecology	AI, peer, teacher, hybrid sequences	Controlled comparative designs	Match source to feedback level and learner need
Equity and policy	Access, privacy, disclosure, discipline	Multi-site and policy-intervention studies	Avoid widening feedback and assessment inequalities

LIMITATIONS

The review's methodological architecture imposes specific interpretive constraints. The reliance on Scopus as the exclusive bibliographic database inherently circumscribes the evidence base. Relevant empirical investigations indexed exclusively within ERIC, Web of Science, LLBA, ProQuest, regional databases, or institutional repositories may therefore be systematically absent from the synthesis. The restriction to English-language publications concurrently risks the underrepresentation of localised scholarship. Temporal coverage for the 2026 publication year terminated on 20 August, rendering the final annual volume necessarily incomplete.

A structural limitation dictates the analytical depth of the synthesis. Because the exported corpus comprised only bibliographic metadata and abstracts rather than full-text manuscripts, the screening and synthesis protocols were fundamentally limited by the information disclosed in these abbreviated fields. Abstracts vary in descriptive precision, often prioritise positive statistical findings, and rarely provide the methodological granularity needed for a comprehensive risk-of-bias assessment. The current study must consequently be interpreted as a systematic abstract-level evidence map rather than a formal meta-analysis or a methodological substitute for duplicate full-text review. The execution of screening procedures by a single reviewer, despite the integration of deterministic assistance and targeted manual verification, introduces potential selection bias at the margins of inclusion.

Analytical metrics present further constraints. Outcome counts utilised non-exclusive categorisation and does not quantify aggregate effect magnitudes. A substantial proportion of the included studies relied upon convenience sampling, self-reported perception metrics, abbreviated intervention timelines, and locally validated psychometric instruments. Aggregated tool labels simultaneously obscure critical version differences within rapidly iterating generative models. This publication velocity generates severe temporal instability. Generative systems, institutional governance policies, and localised student practices evolve faster than conventional academic review cycles.

CONCLUSION

Algorithmic reliance within ESL/EFL higher education operates fundamentally as a complex pattern of cognitive delegation rather than a unidimensional metric of usage frequency. Students consistently direct automated support toward localised grammar correction, vocabulary refinement, structural paraphrasing, initial ideation, organisational planning, diagnostic feedback, and iterative revision. These documented practices yield immediate improvements in textual quality, attenuate writing anxiety, and elevate domain confidence, particularly when algorithmic feedback is delivered rapidly within highly structured instructional environments. High perceived utility does not guarantee educational productivity. Positive

perceptions of technology do not inherently make algorithmic application trustworthy or pedagogically sound.

The synthesised evidence isolates two definitive pedagogical realities. AI systems support language acquisition most reliably when learners actively retain fundamental authorship, systematically verify generated information, critically evaluate machine feedback, and situate the tool within reflective teacher or peer-mediated instructional ecologies. Conversely, generative AI actively risks severe learning displacement when artificial linguistic fluency substitutes for autonomous planning, logical reasoning, source evaluation, and independent conceptual expression. Guided, process-oriented integration thus offers a significantly more defensible pedagogical architecture than either unrestricted algorithmic delegation or blanket institutional prohibition.

Higher education institutions must operationalise calibrated reliance. This objective is achieved by restricting algorithmic delegation to bounded tasks, explicitly preserving learning-critical cognitive labour, rendering verification routines structurally visible, and assessing the longitudinal writing process alongside the final textual product. Subsequent empirical research must transcend subjective attitude surveys to establish objective behavioural measures of systemic dependency, long-term skill transfer, independent unaided performance, and critical authorial ownership. The enduring pedagogical challenge resides not in determining whether students will apply generative writing tools, but in structuring institutional environments that compel learners to interact with these systems without surrendering the essential cognitive and ethical labour through which academic writing develops.

REFERENCES

1. Abdullah, A. M. A., Nasim, S. M., & Alqahtani, M. A. (2026). The role of Grammarly in Saudi EFL writing: Students' perceptions of its effectiveness, efficiency, and overreliance. FWU Journal of Social Sciences, 20(1), 134–148. https://doi.org/10.51709/19951272/Spring2026/10
2. Aksakallı, C., & Daşer, Z. (2025). Unlocking EFL learners' insights into ChatGPT use for L2 writing: The impacts of usage frequency and gender variations. Current Psychology, 44(9), 7957–7977. https://doi.org/10.1007/s12144-025-07437-3
3. Al-Alami, S. E. (2024). EFL learners' attitudes towards utilizing ChatGPT for acquiring writing skills in higher education: A case study of computing students. Journal of Language Teaching and Research, 15(4), 1029–1038. https://doi.org/10.17507/JLTR.1504.01
4. Alkamel, M. A. A., & 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 and Humanities Open, 10, Article 101095. https://doi.org/10.1016/j.ssaho.2024.101095
5. Alnemrat, A., Aldamen, H., Almashour, M., Al-Deaibes, M., & AlSharefeen, R. (2025). AI vs. teacher feedback on EFL argumentative writing: A quantitative study. Frontiers in Education, 10, Article 1614673. https://doi.org/10.3389/feduc.2025.1614673
6. Alshehri, O. A. O., Lamouchi, A., Abdellatif, M. S., Al-Dosari, M. N. A., Mekheimer, M., & Nemt-allah, M. A. (2026). From tool to scaffold: Structured human-AI collaboration and its effects on academic writing and digital critical thinking among Saudi EFL learners. Frontiers in Psychology, 17, 1–12. https://doi.org/10.3389/fpsyg.2026.1830103
7. Alsofyani, A. H., & Barzanji, A. M. (2025). The effects of ChatGPT-generated feedback on Saudi EFL learners' writing skills and perception at the tertiary level: A mixed-methods study. Journal of Educational Computing Research, 63(2), 431–463. https://doi.org/10.1177/07356331241307297
8. Bekturova, M., Tulepova, S., & Zhaitapova, A. (2025). Predicting Kazakhstani TEFL students' continuance intention towards using ChatGPT in academic writing. Education and Information Technologies, 30(9), 11889–11909. https://doi.org/10.1007/s10639-024-13306-6
9. Boudouaia, A., Mouas, S., & Kouider, B. (2024). A study on ChatGPT-4 as an innovative approach to enhancing English as a foreign language writing learning. Journal of Educational Computing Research, 62(6), 1509–1537. https://doi.org/10.1177/07356331241247465

10. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa
11. Cengiz, B. C., Bilki, Z., Ataş, A. H., & Celik, B. (2025). Exploring second language writers' engagement with ChatGPT feedback: Revision behaviors and perceptions. System, 134, Article 103837. https://doi.org/10.1016/j.system.2025.103837
12. Chen, J., Huang, K., Lai, C., & Jin, T. (2025). The impact of GenAI-based collaborative inquiry on critical thinking in argumentation: A case study of blended argumentative writing pedagogy. TESOL Quarterly, 59(S1), S150–S182. https://doi.org/10.1002/tesq.3407
13. Cho, H. (2025). Learner autonomy and anxiety in AI-assisted L2 writing: A PLS-SEM study of QuillBot use. English Teaching (South Korea), 80(4), 85–103. https://doi.org/10.15858/engtea.80.4.202512.85
14. Dahmash, N. B. (2026). "I no longer need a tutor": Saudi EFL students' independent use of ChatGPT for essay writing and grammar learning. Arab World English Journal, 2026-January(3), 147–165. https://doi.org/10.24093/awej/AI3.10
15. Dizon, G., & Gold, J. (2023). Exploring the effects of Grammarly on EFL students' foreign language anxiety and learner autonomy. JALT CALL Journal, 19(3), 299–316. https://doi.org/10.29140/jaltcall.v19n3.1049
16. Dong, L. (2024). "Brave new world" or not?: A mixed-methods study of the relationship between second language writing learners' perceptions of ChatGPT, behaviors of using ChatGPT, and writing proficiency. Current Psychology, 43(21), 19481–19495. https://doi.org/10.1007/s12144-024-05728-9
17. 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, 167–189. https://doi.org/10.32038/ltrq.2024.43.10
18. Estrella, F. (2025). Is Grammarly useful for improving essay-writing skills? Ecuadorian undergraduate polytechnic students' viewpoints. Journal of Applied Research in Higher Education, 18(4), 993–1014. https://doi.org/10.1108/JARHE-10-2024-0529
19. Fan, N. (2023). Exploring the effects of automated written corrective feedback on EFL students' writing quality: A mixed-methods study. SAGE Open, 13(2). https://doi.org/10.1177/21582440231181296
20. Gao, J., Zhang, J., & Li, Y. (2025). Do AI chatbot-integrated writing tasks influence writing self-efficacy and critical thinking ability? An exploratory study. Computers and Education: Artificial Intelligence, 9, Article 100472. https://doi.org/10.1016/j.caeai.2025.100472
21. Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. Internet and Higher Education, 63, Article 100962. https://doi.org/10.1016/j.iheduc.2024.100962
22. Han, F., & Wang, Z. (2026). Teacher feedback and ChatGPT feedback on Chinese university EFL learners' English essay revision: A mixed-methods study. Computers and Education: Artificial Intelligence, 10, Article 100590. https://doi.org/10.1016/j.caeai.2026.100590
23. Hao, H., Razali, A. B., & Zuo, R. (2026). Exploring the impact of integrated AWE and generative AI feedback on Chinese EFL undergraduates' higher-order thinking in argumentative writing. SAGE Open, 16(1). https://doi.org/10.1177/21582440251413884
24. Hong, H., Vate-U-Lan, P., & Viriyavejakul, C. (2025). Cognitive offload instruction with generative AI: A quasi-experimental study on critical thinking gains in English writing. Forum for Linguistic Studies, 7(7), 325–334. https://doi.org/10.30564/fls.v7i7.10072
25. Hong, S., & Shin, Y. K. (2025). Effects of three levels of AI integration on second language academic writing: Evaluating restricted, guided, and free use of ChatGPT. System, 134, Article 103820. https://doi.org/10.1016/j.system.2025.103820
26. Hou, X. L., He, S. Y., & Cui, G. R. X. (2024). Learner use of AI-generated feedback for written corrective feedback in L2 writing: Usefulness, user proficiency and attitude. ACM International Conference Proceeding Series, 70–76. https://doi.org/10.1145/3678726.3678767
27. Hu, X., & Gong, W. (2025). Modeling Chinese EFL learners' intention to use generative AI for L2 writing through an integrated model of the TAM and TTF. Education and Information Technologies, 30(13), 18157–18179. https://doi.org/10.1007/s10639-025-13505-9

28. Hu, Y., Curdt-Christiansen, X. L., & Wang, J. (2025). Smart tools, smarter minds? Learner-AI interaction and AI assistance on critical thinking in EAP contexts. Journal of English for Academic Purposes, 78, Article 101586. https://doi.org/10.1016/j.jeap.2025.101586
29. Jiang, Y., Wu, Q., Yang, Y., Jian, C., & Zhao, J. (2026). Learner agency in revising GenAI-generated statements of purpose. British Journal of Educational Technology, 57(4), 965–983. https://doi.org/10.1111/bjet.70041
30. Jin, F., Lin, C.-H., & Lai, C. (2026). Profiles of ESL postgraduate students' use of artificial intelligence in writing: Antecedents and outcomes. Computers and Education, 253, Article 105689. https://doi.org/10.1016/j.compedu.2026.105689
31. Kailin, Z., & Saeed, M. A. (2026). Chinese EFL learners' engagement with ChatGPT feedback on academic writing: A case study in Malaysia. Computers and Composition, 79, Article 102976. https://doi.org/10.1016/j.compcom.2025.102976
32. Kohnke, L. (2024). Exploring EAP students' perceptions of GenAI and traditional grammar-checking tools for language learning. Computers and Education: Artificial Intelligence, 7, Article 100279. https://doi.org/10.1016/j.caeai.2024.100279
33. Kurt, G., & Kurt, Y. (2024). Enhancing L2 writing skills: ChatGPT as an automated feedback tool. Journal of Information Technology Education: Research, 23, Article 24. https://doi.org/10.28945/5370
34. Lee, Y.-J., Davis, R. O., & Lee, S. O. (2024). University students' perceptions of artificial intelligence-based tools for English writing courses. Online Journal of Communication and Media Technologies, 14(1), Article e202412. https://doi.org/10.30935/ojcmt/14195
35. Li, T., Liu, Z., & Zou, D. (2026). Understanding the usefulness-risk paradox in AI writing tool adoption: An extended technology acceptance model study in Chinese university L2 writing contexts. PLOS One, 21(8), Article e0355494. https://doi.org/10.1371/journal.pone.0355494
36. Li, X., Wang, W., & Gu, L. (2025). An empirical study of Chinese EFL student writers' use of GenAI feedback during revision: Perceptions, engagement and motivational impact. Assessment and Evaluation in Higher Education. https://doi.org/10.1080/02602938.2025.2587925
37. Liang, J., Huang, F., & Teo, T. (2024). Understanding Chinese university EFL learners' perceptions of AI in English writing. International Journal of Computer-Assisted Language Learning and Teaching, 14(1). https://doi.org/10.4018/IJCALLT.358918
38. Lin, Y., & Yu, Z. (2025). Elucidating university students' intentions to seek automated writing feedback from Grammarly: Toward perceptual and systemic predictors. Humanities and Social Sciences Communications, 12(1), Article 7. https://doi.org/10.1057/s41599-024-03861-1
39. Liu, C., Hou, J., Tu, Y.-F., Wang, Y., & Hwang, G.-J. (2023). Incorporating a reflective thinking promoting mechanism into artificial intelligence-supported English writing environments. Interactive Learning Environments, 31(9), 5614–5632. https://doi.org/10.1080/10494820.2021.2012812
40. Lu, D., & Zeng, Y. (2025). Exploring the use of GenAI in model texts as a feedback instrument: L2 students' practice and perceptions. Language Teaching Research. https://doi.org/10.1177/13621688251397578
41. Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. Smart Learning Environments, 11(1), Article 9. https://doi.org/10.1186/s40561-024-00295-9
42. Mat, A. C., Zulkornain, L. H., & Rahman, N. A. A. (2024). Automated writing evaluation: Users' perception and expectations. International Journal of Information and Education Technology, 14(2), 183–192. https://doi.org/10.18178/ijiet.2024.14.2.2039
43. Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: A study on proficiency, revision frequency and writing quality. Discover Education, 4(1), Article 170. https://doi.org/10.1007/s44217-025-00602-7
44. Miranty, D., Widiati, U., Cahyono, B. Y., & Sharif, T. I. S. T. (2023). Automated writing evaluation tools for Indonesian undergraduate English as a foreign language students' writing. International Journal of Evaluation and Research in Education, 12(3), 1705–1715. https://doi.org/10.11591/ijere.v12i3.24958

45. Mohammad, T., Nazim, M., Alzubi, A. A. F., & Khan, S. I. (2024). Examining EFL students' motivation level in using QuillBot to improve paraphrasing skills. World Journal of English Language, 14(1), 501–510. https://doi.org/10.5430/wjel.v14n1p501
46. Nelson, A. S., Santamaría, P. V., Javens, J. S., & Ricaurte, M. (2025). Students' perceptions of generative artificial intelligence (GenAI) use in academic writing in English as a foreign language. Education Sciences, 15(5), Article 611. https://doi.org/10.3390/educsci15050611
47. Nguyen, L. Q., Le, H. V., & Nguyen, P. T. (2025). A mixed-methods study on the use of ChatGPT in the pre-writing stage: EFL learners' utilization patterns, affective engagement, and writing performance. Education and Information Technologies, 30(8), 10511–10534. https://doi.org/10.1007/s10639-024-13231-8
48. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, Article n71. https://doi.org/10.1136/bmj.n71
49. Pinninti, L. R. (2024). Undergraduate ESL students' use and perceptions of ChatGPT for academic writing purposes [O uso e as percepções de estudantes de graduação em ESL sobre o ChatGPT para fins de escrita acadêmica]. Texto Livre, 18, Article e58320. https://doi.org/10.1590/1983-3652.2025.58320
50. Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. Cogent Education, 11(1), Article 2410101. https://doi.org/10.1080/2331186X.2024.2410101
51. Ruyang, L., Zhang, C., & Ye, H. (2026). When does GenAI feedback support learning? Trust calibration and verification in L2 writing assessment. Assessing Writing, 69, Article 101087. https://doi.org/10.1016/j.asw.2026.101087
52. Salam, U. (2025). The integration of ChatGPT in English for foreign language course: Elevating AI writing assistant acceptance. Computers in the Schools, 42(2), 145–165. https://doi.org/10.1080/07380569.2024.2446239
53. Shen, Y., & Chen, L. (2025). 'Critical chatting' or 'casual cheating': How graduate EFL students utilize ChatGPT for academic writing. Computer Assisted Language Learning. https://doi.org/10.1080/09588221.2025.2479141
54. Shi, H., Chai, C. S., Zhou, S., & Aubrey, S. (2026). Comparing the effects of ChatGPT and automated writing evaluation on students' writing and ideal L2 writing self. Computer Assisted Language Learning, 39(4), 1148–1175. https://doi.org/10.1080/09588221.2025.2454541
55. Tai, H.-Y., Lin, M.-F., & Chen, Y.-S. (2026). Incorporating ChatGPT into genre-based instruction for argumentative writing among EFL college students. International Journal of Applied Linguistics, 36(1), 767–781. https://doi.org/10.1111/ijal.12777
56. Tam, A. C. F. (2025). Interacting with ChatGPT for internal feedback and factors affecting feedback quality. Assessment and Evaluation in Higher Education, 50(2), 219–235. https://doi.org/10.1080/02602938.2024.2374485
57. Tanharat, N., & Phootirat, P. (2025). Integrating generative AI in EFL academic writing: Thai English-major students' purposes, perceptions, and experiences with ChatGPT. rEFLections, 32(3), 1891–1916. https://doi.org/10.61508/refl.v32i3.285982
58. Taskiran, A., Yazici, M., & Erdem, A. I. (2024). Contribution of automated feedback to the English writing competence of distance foreign language learners. E-Learning and Digital Media, 21(1), 24–41. https://doi.org/10.1177/20427530221139579
59. Tekir, S. (2026). Generative AI use in EFL writing: Associations with originality, critical reasoning, and metacognitive engagement in a Turkish higher education context. Computer Assisted Language Learning. https://doi.org/10.1080/09588221.2026.2617399
60. Teng, M. F. (2025). Metacognitive awareness and EFL learners' perceptions and experiences in utilising ChatGPT for writing feedback. European Journal of Education, 60(1), Article e12811. https://doi.org/10.1111/ejed.12811

61. Thi, N. K., Nikolov, M., & Simon, K. (2023). Higher-proficiency students' engagement with and uptake of teacher and Grammarly feedback in an EFL writing course. Innovation in Language Learning and Teaching, 17(3), 690–705. https://doi.org/10.1080/17501229.2022.2122476
62. Thi, T. T. M. (2024). Understanding Vietnamese English majors' use and acceptance of AI-powered tools for English academic writing at university. Taiwan International ESP Journal, 15(2), 29–54. https://doi.org/10.6706/TIESPJ.202412_15(2).0002
63. Wang, Y. (2024). Cognitive and sociocultural dynamics of self-regulated use of machine translation and generative AI tools in academic EFL writing. System, 126, Article 103505. https://doi.org/10.1016/j.system.2024.103505
64. Werdiningsih, I., Marzuki, & Rusdin, D. (2024a). Balancing AI and authenticity: EFL students' experiences with ChatGPT in academic writing. Cogent Arts and Humanities, 11(1), Article 2392388. https://doi.org/10.1080/23311983.2024.2392388
65. Werdiningsih, I., Marzuki, Indrawati, I., Rusdin, D., Ivone, F. M., Basthomi, Y., & Zulfahreza. (2024b). Revolutionizing EFL writing: Unveiling the strategic use of ChatGPT by Indonesian master's students. Cogent Education, 11(1), Article 2399431. https://doi.org/10.1080/2331186X.2024.2399431
66. Wu, C., Moorhouse, B. L., Wan, Y., & Wu, M. (2026). Exploring PhD students' utilization of generative AI in academic writing for publication purposes: Insights for EAP. Journal of English for Academic Purposes, 79, Article 101612. https://doi.org/10.1016/j.jeap.2025.101612
67. Wulyani, A. N., Widiati, U., Muniroh, S., Rachmadhany, C. D., Nurlaila, N., Hanifiyah, L., & Sharif, T. I. S. T. (2024). Patterns of utilizing AI-assisted tools among EFL students: Need surveys for assessment model development. LLT Journal: Journal on Language and Language Teaching, 27(1), 157–173. https://doi.org/10.24071/llt.v27i1.7966
68. Xu, T., & Jumaat, N. F. (2024). ChatGPT-empowered writing strategies in EFL students' academic writing: Calibre, challenges and chances. International Journal of Interactive Mobile Technologies, 18(15), 95–114. https://doi.org/10.3991/ijim.v18i15.49219
69. Yang, H., Gao, C., & Shen, H.-Z. (2024). Learner interaction with, and response to, AI-programmed automated writing evaluation feedback in EFL writing: An exploratory study. Education and Information Technologies, 29(4), 3837–3858. https://doi.org/10.1007/s10639-023-11991-3
70. Yang, W., & Lin, C. (2025). Translanguaging with generative AI in EFL writing: Students' practices and perceptions. Journal of Second Language Writing, 67, Article 101181. https://doi.org/10.1016/j.jslw.2025.101181
71. Yang, Y. (2025). Exploring factors influencing L2 learners' use of GAI-assisted writing technology: Based on the UTAUT model. Asia Pacific Journal of Education. https://doi.org/10.1080/02188791.2025.2505664
72. Yeung, S. (2025). University students' engagement with generative AI-supported automated writing evaluation (AWE) feedback. Journal of Second Language Writing, 68, Article 101203. https://doi.org/10.1016/j.jslw.2025.101203
73. Yildiz, H., & Gonen, S. I. K. (2024). Automated writing evaluation system for feedback in the digital world: An online learning opportunity for English as a foreign language students. Turkish Online Journal of Distance Education, 25(3), 184–206. https://doi.org/10.17718/TOJDE.1169727
74. Yin, X., & Dou, K. (2025). An AI-assisted critical thinking intervention to enhance undergraduate EFL learners' writing proficiency. Studies in Educational Evaluation, 86, Article 101480. https://doi.org/10.1016/j.stueduc.2025.101480
75. Zhang, J., & Wang, J. (2025). Student perceptions of hybrid feedback: Using Gen-AI to enhance engagement with EAP writing feedback. JALT CALL Journal, 21(2), Article 2175. https://doi.org/10.29140/jaltcall.v21n2.2175
76. Zhang, Z., & Hyland, K. (2025). The role of digital literacy in student engagement with automated writing evaluation (AWE) feedback on second language writing. Computer Assisted Language Learning, 38(5-6), 1060–1085. https://doi.org/10.1080/09588221.2023.2256815
77. Zhang, Z., Wang, C., & Shang, Y. (2026). Are AI tools effective in reducing cognitive load or boosting writing self-efficacy for L2 postgraduates? Asia-Pacific Education Researcher, 35(2), 343–354. https://doi.org/10.1007/s40299-025-01031-8

78. Zou, S., Guo, K., Wang, J., & Liu, Y. (2026). Investigating students' uptake of teacher- and ChatGPT-generated feedback in EFL writing: A comparison study. Computer Assisted Language Learning, 39(4), 1062–1091. https://doi.org/10.1080/09588221.2024.2447279

APPENDIX A: COMPLETE SCOPUS SEARCH STRING

```
TITLE-ABS-KEY(
  (
    "AI writing tool*" OR "AI writing assistant*" OR
    "artificial intelligence writing tool*" OR
    "artificial intelligence writing assistant*" OR
    "AI-assisted writing" OR "AI-supported writing" OR
    "AI-generated text" OR "machine-generated text" OR
    "generative artificial intelligence" OR "generative AI" OR GenAI OR
    "large language model*" OR LLM* OR ChatGPT OR "GPT-3" OR "GPT-4" OR
    "GPT-4o" OR Grammarly OR QuillBot OR "Google Bard" OR "Google Gemini" OR
    "Microsoft Copilot" OR "Claude AI" OR "automated writing evaluation" OR
    "automated writing feedback" OR "automated essay scoring"
  )
  AND
  (
    EFL OR ESL OR "English as a foreign language" OR
    "English as a second language" OR "English language learner*" OR
    "second language learner*" OR "foreign language learner*" OR
    "second language writing" OR "L2 writing" OR "foreign language writing" OR
    "English for academic purposes" OR EAP OR "academic English"
  )
  AND (student* OR learner*)
  AND
  ("higher education" OR universit* OR college* OR undergraduate* OR
   postgraduate* OR "graduate student*" OR "tertiary education")
```

```
AND
(reliance OR reliant OR rely OR dependenc* OR overreliance OR "over-reliance" OR
 trust OR adopt* OR accept* OR usage OR utilisation OR utilization OR
 "frequency of use" OR "purpose of use" OR attitude* OR perception* OR
 intention* OR "writing skill*" OR "writing ability" OR "writing competenc*" OR
 "writing performance" OR "writing quality" OR "critical thinking" OR
 "cognitive offloading")
)
AND PUBYEAR > 2022 AND PUBYEAR < 2027
AND (LIMIT-TO(DOCTYPE, "ar") OR LIMIT-TO(DOCTYPE, "cp") OR LIMIT-TO(DOCTYPE, "re"))
AND LIMIT-TO(LANGUAGE, "English")
```

APPENDIX B: DETAILED ELIGIBILITY RULES

Criterion	Include	Exclude
Population	Higher-education students: undergraduate, postgraduate, doctoral, college, university, or tertiary learners	School-level learners; teacher-only or supervisor-only samples
Context	ESL, EFL, L2 English writing, EAP, academic English, or explicit second-language writing in EMI	No ESL/EFL/L2 writing dimension
Concept	Learner-facing AI writing support for ideation, drafting, language, paraphrasing, feedback, or revision	AI detection, materials development, model scoring, or incidental tool mention
Outcomes	Use/purpose, acceptance, trust, dependency, overreliance, writing, critical thinking, autonomy, or self-regulation	No relevant reliance, writing, or thinking outcome
Evidence	Empirical article or conference paper with an identifiable design/sample in the abstract	Review, commentary, conceptual paper, or no empirical evidence
Limits	English; 2023 through 20 August 2026	Other language or outside date range

Note. Reviews were retained for contextual orientation but excluded from the primary empirical synthesis

APPENDIX C: CORE EVIDENCE MATRIX

The core evidence matrix below expands Table 2 with design, tool, outcomes, and DOI information for studies most directly informing the synthesis. It is a representative analytical matrix drawn from the complete 298-record coded corpus.

Study	Tool	Design	Coded outcomes	DOI
Estrella, 2025	Grammarly	Mixed methods	Frequency/purpose/practices; Acceptance/perceptions; Writing performance/skills	10.1108/JARHE-10-2024-0529
Cengiz et al.,	ChatGPT/GPT;	Mixed methods	Frequency/purpose/practice	10.1016/j.system.2025.103837

2025	General GenAI/LLM		s; Acceptance/perceptions; Writing performance/skills	
Hu et al., 2025	Other/unspecified AI writing tool	Mixed methods	Frequency/purpose/practices; Acceptance/perceptions; Critical thinking/reasoning; Autonomy/self-regulation	10.1016/j.jeap.2025.101586
Li et al., 2026	General GenAI/LLM	Survey/correlational	Frequency/purpose/practices; Acceptance/perceptions; Trust/dependency/overreliance; Writing performance/skills	10.1371/journal.pone.0355494
Teng, 2025	ChatGPT/GPT	Mixed methods	Frequency/purpose/practices; Acceptance/perceptions; Critical thinking/reasoning; Autonomy/self-regulation	10.1111/ejed.12811
Chen et al., 2025	General GenAI/LLM	Longitudinal/case/action research	Frequency/purpose/practices; Critical thinking/reasoning	10.1002/tesq.3407
Tanharat & Phootirat, 2025	ChatGPT/GPT; General GenAI/LLM	Mixed methods	Frequency/purpose/practices; Acceptance/perceptions; Writing performance/skills; Autonomy/self-regulation	10.61508/refl.v32i3.285982
Aksakallı & Daşer, 2025	ChatGPT/GPT	Mixed methods	Frequency/purpose/practices; Acceptance/perceptions; Trust/dependency/overreliance	10.1007/s12144-025-07437-3
Dahmash, 2026	ChatGPT/GPT	Qualitative	Frequency/purpose/practices; Acceptance/perceptions; Trust/dependency/overreliance; Autonomy/self-regulation	10.24093/awej/AI3.10
Mekheimer, 2025	Grammarly; General GenAI/LLM	Experimental/quasi-experimental	Frequency/purpose/practices; Trust/dependency/overreliance; Writing performance/skills; Critical thinking/reasoning	10.1007/s44217-025-00602-7
Abdullah et al., 2026	Grammarly; AWE/AWCF	Survey/correlational	Frequency/purpose/practices; Acceptance/perceptions; Trust/dependency/overreliance; Writing performance/skills; Autonomy/self-regulation	10.51709/19951272/Spring2026/10
Wu et al., 2026	General GenAI/LLM	Qualitative	Frequency/purpose/practices; Acceptance/perceptions	10.1016/j.jeap.2025.101612
Alnemrat et al., 2025	General GenAI/LLM	Experimental/quasi-experimental	Writing performance/skills	10.3389/feduc.2025.1614673
Tekir, 2026	ChatGPT/GPT; Grammarly; QuillBot; General GenAI/LLM	Mixed methods	Frequency/purpose/practices; Trust/dependency/overreliance; Writing performance/skills; Critical thinking/reasoning; Autonomy/self-regulation	10.1080/09588221.2026.2617399
Nguyen et al., 2025	ChatGPT/GPT	Mixed methods	Frequency/purpose/practices; Acceptance/perceptions; Writing performance/skills	10.1007/s10639-024-13231-8
Ruyang et al.,	General	Mixed methods	Frequency/purpose/practice	10.1016/j.asw.2026.101087

2026	GenAI/LLM		s; Acceptance/perceptions; Trust/dependency/overreliance; Writing performance/skills	
Hong et al., 2025	General GenAI/LLM	Experimental/quasi-experimental	Writing performance/skills; Critical thinking/reasoning; Autonomy/self-regulation	10.30564/fls.v7i7.10072
Zou et al., 2026	ChatGPT/GPT; AWE/AWCF	Survey/correlational	Frequency/purpose/practices; Acceptance/perceptions; Writing performance/skills	10.1080/09588221.2024.2447279
Shen & Chen, 2025	ChatGPT/GPT; General GenAI/LLM	Longitudinal/case/action research	Frequency/purpose/practices; Acceptance/perceptions; Critical thinking/reasoning	10.1080/09588221.2025.2479141
Jin et al., 2026	ChatGPT/GPT	Survey/correlational	Frequency/purpose/practices; Acceptance/perceptions; Critical thinking/reasoning	10.1016/j.compedu.2026.105689
Hao et al., 2026	ChatGPT/GPT; AWE/AWCF; General GenAI/LLM	Experimental/quasi-experimental	Frequency/purpose/practices; Trust/dependency/overreliance; Writing performance/skills; Critical thinking/reasoning	10.1177/21582440251413884
Gao et al., 2025	General GenAI/LLM	Experimental/quasi-experimental	Critical thinking/reasoning	10.1016/j.caeai.2025.100472
Han & Wang, 2026	ChatGPT/GPT	Mixed methods	Acceptance/perceptions; Writing performance/skills	10.1016/j.caeai.2026.100590
Alshehri et al., 2026	General GenAI/LLM	Mixed methods	Frequency/purpose/practices; Writing performance/skills; Critical thinking/reasoning; Autonomy/self-regulation	10.3389/fpsyg.2026.1830103
Yin & Dou, 2025	ChatGPT/GPT	Experimental/quasi-experimental	Frequency/purpose/practices; Acceptance/perceptions; Trust/dependency/overreliance; Writing performance/skills; Critical thinking/reasoning	10.1016/j.stueduc.2025.101480
Li et al., 2025	General GenAI/LLM	Mixed methods	Frequency/purpose/practices; Acceptance/perceptions; Writing performance/skills	10.1080/02602938.2025.2587925
Bekturova et al., 2025	ChatGPT/GPT	Qualitative	Frequency/purpose/practices; Acceptance/perceptions; Writing performance/skills	10.1007/s10639-024-13306-6
Hong & Shin, 2025	ChatGPT/GPT; General GenAI/LLM	Qualitative	Frequency/purpose/practices; Writing performance/skills; Autonomy/self-regulation	10.1016/j.system.2025.103820
El-Garawany, 2024	QuillBot	Experimental/quasi-experimental	Acceptance/perceptions; Writing performance/skills	10.32038/ltrq.2024.43.10
Werdiningsih et al., 2024a	ChatGPT/GPT	Longitudinal/case/action research	Trust/dependency/overreliance; Writing performance/skills; Critical thinking/reasoning	10.1080/23311983.2024.2392388
Mahapatra, 2024	ChatGPT/GPT; AWE/AWCF; General GenAI/LLM	Mixed methods	Acceptance/perceptions; Writing performance/skills	10.1186/s40561-024-00295-9
Thi et al.,	Grammarly;	Empirical design	Trust/dependency/overrelia	10.1080/17501229.2022.2122

2023	AWE/AWCF	not fully specified in abstract	nce; Writing performance/skills	476
Liu et al., 2023	AWE/AWCF	Experimental/quasi-experimental	Acceptance/perceptions; Writing performance/skills; Autonomy/self-regulation	10.1080/10494820.2021.2012812
Dizon & Gold, 2023	Grammarly; AWE/AWCF	Qualitative	Acceptance/perceptions; Writing performance/skills; Autonomy/self-regulation	10.29140/jaltcall.v19n3.1049
Pinninti, 2024	ChatGPT/GPT	Survey/correlational	Frequency/purpose/practices; Acceptance/perceptions; Trust/dependency/overreliance	10.1590/1983-3652.2025.58320
Guo et al., 2024	ChatGPT/GPT	Experimental/quasi-experimental	Writing performance/skills	10.1016/j.iheduc.2024.100962
Wulyani et al., 2024	Grammarly	Survey/correlational	Frequency/purpose/practices; Acceptance/perceptions	10.24071/llt.v27i1.7966
Wang, 2024	ChatGPT/GPT; General GenAI/LLM	Mixed methods	Acceptance/perceptions; Writing performance/skills; Autonomy/self-regulation	10.1016/j.system.2024.103505
Werdiningsih et al., 2024b	ChatGPT/GPT	Mixed methods	Frequency/purpose/practices; Trust/dependency/overreliance; Writing performance/skills	10.1080/2331186X.2024.2399431
Dong, 2024	ChatGPT/GPT	Mixed methods	Acceptance/perceptions; Writing performance/skills; Autonomy/self-regulation	10.1007/s12144-024-05728-9
Polakova & Ivenz, 2024	ChatGPT/GPT	Mixed methods	Frequency/purpose/practices; Acceptance/perceptions; Writing performance/skills	10.1080/2331186X.2024.2410101
Fan, 2023	Grammarly; AWE/AWCF	Mixed methods	Acceptance/perceptions; Writing performance/skills	10.1177/21582440231181296
Kohnke, 2024	ChatGPT/GPT; Grammarly; General GenAI/LLM	Qualitative	Frequency/purpose/practices; Acceptance/perceptions; Trust/dependency/overreliance	10.1016/j.caeai.2024.100279
Xu & Jumaat, 2024	ChatGPT/GPT	Mixed methods	Frequency/purpose/practices; Critical thinking/reasoning	10.3991/ijim.v18i15.49219
Thi, 2024	General GenAI/LLM	Qualitative	Frequency/purpose/practices; Acceptance/perceptions; Trust/dependency/overreliance; Critical thinking/reasoning	10.6706/TIESPJ.202412_15(2).0002

Note. This matrix contains the studies most directly used in the narrative synthesis; coding categories were non-exclusive.