



Systematic Literature Review

Generative AI-Assisted Writing in Indonesian Language Education: A Systematic Literature Review of Process Pedagogy, Critical Literacy, and Academic Integrity

Zathu Restie Utami

Universitas Islam Negeri Raden Intan Lampung, Indonesia

Article Info

Article history:

Received 16 July 2026

Revised 19 September 2026

Accepted 21 September 2026

Keywords:

Generative artificial intelligence

Indonesian language education

Writing pedagogy

Critical AI literacy

Academic integrity



Abstract

Generative artificial intelligence (GenAI) has rapidly entered writing classrooms, yet its educational value depends on whether it supports or bypasses the intellectual work of composing. This systematic literature review synthesizes 20 peer-reviewed international journal articles published between 2022 and 2025 to identify how GenAI, educational chatbots, digital multimodal composing, and critical digital literacy can inform Indonesian language education. A transparent search and screening protocol was applied to journal articles concerned with language learning, writing development, feedback, multimodal composition, authorship, equity, and academic integrity. The thematic synthesis produced five findings. First, GenAI can assist brainstorming, genre modelling, language reformulation, and formative feedback, particularly when students remain responsible for planning and revision. Second, process-based designs are more educationally defensible than product substitution because they make decision making and authorship visible. Third, hallucination, bias, privacy, overdependence, and unreliable AI-detection systems require explicit critical AI literacy. Fourth, multimodal and narrative research indicates that technology should strengthen meaning, cultural voice, and audience awareness rather than reward surface polish. Fifth, equitable implementation requires flexible access, teacher mediation, and assessment that values evidence, reflection, and revision history. Based on these findings, the article proposes a five-stage HUMAN-AI writing cycle for Indonesian language classrooms: human grounding, purposeful prompting, evidence checking, authorial reconstruction, and reflective disclosure. The framework positions GenAI as a supervised dialogic resource, not a substitute author.

Corresponding Author: Zathu Restie Utami, zathurestieutami@radenintan.ac.id

1. Introduction

Writing in language education is not merely the production of grammatically acceptable sentences. It is a cognitive, rhetorical, cultural, and ethical practice through which learners select evidence, develop a position, construct a voice, and anticipate an audience. This broader understanding is especially important in Indonesian language education, where students encounter narrative, exposition, argumentation, review, procedure, and multimodal texts across school and university contexts. The arrival of generative artificial intelligence (GenAI) therefore raises a foundational pedagogical question: does the technology extend the learner's capacity to think through language, or does it remove the very struggle through which writing competence is formed?

Large language models can generate outlines, paragraphs, summaries, explanations, examples, and feedback within seconds. Educational reviews describe substantial possibilities for personalized support, rapid content generation, and interactive assistance, while also emphasizing bias, factual instability, and the need for sustained human oversight (Kasneci et al., 2023; Tlili et al., 2023). In language teaching specifically, ChatGPT and related systems can provide vocabulary support, reformulation, conversational practice, and immediate responses, but teachers and learners need new digital competencies to judge when the output is useful, misleading, culturally inappropriate, or pedagogically counterproductive (Kohnke et al., 2023).

Early empirical evidence is promising but not conclusive. Mahapatra (2024), for example, found that structured use of ChatGPT as a formative feedback tool was associated with improvement in ESL academic writing. Student-perception research similarly reports benefits related to efficiency, explanation, idea development, and confidence, alongside concerns about accuracy, dependency, privacy, and reduced critical engagement (Chan & Hu, 2023; Mogavi et al., 2024). These mixed findings suggest that the central variable is not access to GenAI itself, but the instructional design surrounding its use.

The distinction between assistance and substitution is particularly important for writing. When a learner uses GenAI to test an outline, request counterarguments, compare possible structures, or receive feedback on a self-produced draft, the system may operate as a dialogic scaffold. When a learner delegates planning, drafting, evidence selection, and revision to the system, the final text may conceal the absence of learning. This tension is reflected in academic-integrity discussions, which warn that conventional take-home writing tasks can become unreliable indicators of individual competence unless teachers redesign assessment around process evidence, oral explanation, situated knowledge, and transparent disclosure (Cotton et al., 2024).

A second concern is fairness. AI-output detectors are sometimes treated as objective evidence even though their reliability is contested. Liang et al. (2023) demonstrated that widely used detectors can misclassify non-native English writing as machine-generated. Although the study examined English, the broader principle matters for Indonesian language education: automated judgments should not replace documented pedagogical evidence, teacher knowledge of student development, and due process. In multilingual and linguistically diverse classrooms, simplistic detection can penalize patterned, formulaic, or emerging academic writing that is entirely human.

A third concern is the preservation of cultural and authorial voice. Research on digital storytelling and digital multimodal composing shows that technology can strengthen writing when it supports narrative development, audience awareness, collaboration, and purposeful relationships among words, images, sound, and pacing (Ajabshir, 2024; Cigerci & Yildirim, 2023; Jiang & Hafner, 2024; Yu & Wang, 2025). These studies offer an important counterweight to tool-centered innovation. They imply that the quality of digital writing should be judged by meaning, coherence, voice, and ethical representation, not by technical spectacle or linguistic smoothness alone.

This issue is highly relevant to Indonesian language classrooms because GenAI systems are trained on uneven multilingual data and may reproduce generic diction, dominant cultural patterns, or inaccurate representations of local knowledge. If students accept fluent output without evaluation, the technology can flatten regional experience, obscure source origins, and normalize a detached academic style. Critical digital literacy therefore needs to be expanded into critical AI literacy: the ability to understand the limits of generated text, interrogate evidence, identify bias, protect privacy, document assistance, and retain responsibility for the final communicative act (Darvin & Hafner, 2022; Kim et al., 2025; Zhang & Dikilitaş, 2025).

Existing scholarship is growing rapidly, but it remains fragmented across educational technology, language learning, writing studies, digital storytelling, and AI ethics. Some articles focus on broad educational opportunities and risks; others examine writing feedback, chatbot interaction, multimodal composition, or academic integrity. Fewer studies integrate these strands into a practical framework for Indonesian language education. A synthesis is therefore needed that treats GenAI neither as an automatic solution nor as a threat to be excluded, but as a contested literacy environment requiring purposeful pedagogy.

This review addresses three questions. First, what pedagogical affordances of GenAI and related digital-writing practices are supported by recent international journal literature? Second, what risks, ethical issues, and equity concerns must be addressed before these tools are used in writing assessment? Third, what instructional principles can translate the evidence into a classroom-ready framework for Indonesian

language education? The article contributes a systematic synthesis of 20 recent journal articles and proposes a process-oriented model that foregrounds human agency, evidence checking, authorial reconstruction, and transparent reflection.

2. Method

This study used a systematic literature review with a qualitative thematic-synthesis approach. The protocol was designed to identify recent peer-reviewed journal research that could inform the responsible use of GenAI in Indonesian language writing instruction. Because the topic crosses several research traditions, the review combined literature on generative AI, educational chatbots, second-language writing, digital storytelling, multimodal composing, critical digital literacy, equity, and academic integrity.

The publication window was January 2022 to December 2025. Search terms were organized in three concept groups: (1) technology terms, including 'generative AI,' 'ChatGPT,' 'large language model,' 'educational chatbot,' and 'AI-assisted writing'; (2) language and writing terms, including 'language teaching,' 'writing feedback,' 'digital storytelling,' 'multimodal composing,' and 'literacy'; and (3) governance terms, including 'authorship,' 'ethics,' 'bias,' 'academic integrity,' and 'AI detection.' Searches were conducted through scholarly search engines, DOI records, and publisher journal platforms, followed by backward reference checking from highly relevant review articles.

To improve reproducibility, the search procedure was documented through a structured review log. Searches were conducted between January and February 2026 using Scopus, Web of Science Core Collection, ERIC, Dimensions, Google Scholar, and publisher databases where DOI records were available. The search strategy combined three groups of terms: ("generative AI" OR ChatGPT OR "large language model" OR "AI-assisted writing") AND ("writing education" OR "language learning" OR "digital storytelling" OR "multimodal composing") AND ("academic integrity" OR authorship OR bias OR ethics OR assessment). Google Scholar was used as a supplementary discovery source, with the first 10 result pages screened because of its ranking-based retrieval structure.

The selection process followed a PRISMA-oriented documentation approach. The initial search identified 214 records. After removing 38 duplicates, 176 titles and abstracts were screened. A total of 42 full-text articles were assessed, and 22 were excluded because they did not meet the inclusion criteria: unrelated educational context (9 articles), insufficient methodological or conceptual detail (6 articles), non-peer-reviewed publication type (4 articles), and technical model evaluation without educational relevance (3 articles). The final corpus consisted of 20 journal articles.

Coding was conducted by two reviewers independently. Initial coding disagreements were discussed until consensus was reached, and a third academic reviewer was consulted when interpretation remained uncertain. The coding audit focused on consistency between extracted evidence, thematic categories, and final interpretations. Because the review aimed at qualitative synthesis rather than meta-analysis, no statistical inter-rater reliability coefficient was calculated.

Articles were included when they met four criteria: they were published in a peer-reviewed international journal; they appeared within the review window; they directly addressed education, language learning, writing, multimodal composition, or the ethical governance of generated text; and they provided sufficient conceptual or methodological detail for thematic interpretation. Preprints, book chapters, editorials without analytical development, policy documents, and purely technical model-evaluation papers were excluded. The final corpus contained 20 journal articles.

Each article was coded for research context, educational level, method, tool or digital practice, reported affordances, reported risks, implications for teacher mediation, and relevance to Indonesian language education. Initial descriptive coding was followed by focused thematic comparison. Similar codes were grouped into five domains: writing-process support; feedback and revision; authorship and critical AI literacy; multimodal and cultural meaning; and equity, assessment, and implementation.

The included studies showed heterogeneous evidence characteristics. Among the 20 articles, 8 were empirical studies involving learners or teachers, 7 were conceptual or review-based studies, and 5 examined technology-supported writing practices through qualitative or analytical approaches. The studies covered higher education, teacher education, and language-learning contexts across several countries, with a large

proportion originating from English as an additional language setting. Therefore, conclusions regarding Indonesian language education are interpreted as pedagogical transfer rather than direct empirical evidence.

The synthesis does not claim that findings from English as a second or foreign language transfer automatically to Bahasa Indonesia. Instead, transfer was treated as an analytic inference. Findings were considered applicable when they concerned general writing processes, feedback principles, source evaluation, authorship, multimodal meaning, or classroom governance. Language-specific effects, including the quality of Indonesian output and sensitivity to regional varieties, were treated as areas requiring direct empirical research rather than assumed outcomes.

3. Results and Discussion

3.1 Profile of the evidence and a central pedagogical pattern

The 20 studies represent four overlapping evidence clusters. The first cluster examines GenAI and educational chatbots at a system level, emphasizing opportunities, user perceptions, and ethical concerns (Kasneci et al., 2023; Tlili et al., 2023; Chan & Hu, 2023; Mogavi et al., 2024; Kuhail et al., 2023). The second cluster focuses on language teaching and writing support, including feedback, interaction, and digital competence (Huang et al., 2022; Kohnke et al., 2023; Mahapatra, 2024). The third cluster examines digital storytelling and multimodal composition as established forms of technology-mediated writing pedagogy (Ajabshir, 2024; Cigerci & Yıldırım, 2023; Jiang & Hafner, 2024; Veyis et al., 2025; Yu & Wang, 2025). The fourth cluster addresses critical literacy, AI ethics, equity, authorship, and academic integrity (Darvin & Hafner, 2022; Cotton et al., 2024; Liang et al., 2023; Kim et al., 2025; Rapanta et al., 2025; Zhang & Dikilitaş, 2025). Mukul and Büyüközkan (2023) provide a wider account of educational digital transformation that connects these classroom-level developments to institutional change.

Across these clusters, one pattern is consistent: technology produces educational value when it is embedded in a visible learning process. Chatbots can make responses faster, but speed is not equivalent to learning. Digital storytelling can make texts more engaging, but visual polish is not equivalent to narrative competence. GenAI can generate coherent prose, but coherence is not equivalent to evidence-based understanding. The strongest designs therefore retain human planning, task interpretation, source judgment, revision, and reflection as assessable activities.

This process principle is important because generated text often looks finished before the learner has done the intellectual work normally required to produce it. A conventional writing assignment evaluates both product and process indirectly. GenAI breaks that assumption. Assessment must consequently make the process explicit through outlines, source notes, draft comparisons, prompt records, revision rationales, peer feedback, or brief oral defense. These artifacts do not merely police misconduct; they provide richer evidence of how students develop ideas and make rhetorical decisions.

3.2 GenAI as a scaffold for planning, language support, and formative feedback

Across the reviewed studies, dialogic scaffolding emerged as one frequently reported affordance of GenAI use. Students can ask for examples of text structures, alternative thesis statements, questions that expose missing information, or feedback organized by a teacher-provided rubric. Kohnke et al. (2023) describe language-learning affordances that include interaction, explanation, and individualized support. Mahapatra (2024) provides empirical evidence that ChatGPT can function as a formative feedback tool when students are trained to use it within writing instruction. These findings support a model in which the learner begins with a communicative purpose and a self-produced draft, while AI feedback becomes one source of response rather than the origin of the text.

For Indonesian language education, planning support could be used to compare the organization of exposition, argument, review, or narrative texts. A student might ask the system to identify whether an outline contains a clear claim, whether supporting reasons are logically ordered, or what counterargument has not been considered. The educational goal would not be to copy an improved outline, but to evaluate suggestions against the task, available evidence, audience, and intended voice. Teacher prompts should therefore require choices and justifications, not merely request a completed text.

Language-level support can also be useful, especially for learners who struggle with sentence boundaries, cohesion, paragraph transitions, or formal register. However, surface correction must not be confused with

authorship. A generated revision may alter emphasis, certainty, interpersonal stance, or cultural nuance even when the request appears limited to grammar. Students need to compare versions and identify which changes preserve meaning. This comparison turns AI output into an object of linguistic analysis and can strengthen metalinguistic awareness.

Immediate feedback is attractive in large classes where teachers cannot comment extensively on every draft. Yet the literature does not justify unsupervised replacement of teacher feedback. GenAI tends to produce generic praise, overconfident explanations, or advice disconnected from local assessment expectations. A more reliable design uses teacher-created criteria and bounded prompts, asks students to verify each suggestion, and reserves higher-order decisions for human dialogue. The system may help notice patterns, but the teacher remains responsible for curriculum alignment and the student remains responsible for revision.

Research on chatbot-supported language learning before the widespread release of GenAI also offers a caution. Huang et al. (2022) and Kuhail et al. (2023) found that educational chatbots can support interaction and engagement, but outcomes vary with design quality, learner characteristics, and the instructional role assigned to the system. The novelty of a conversational interface does not guarantee durable learning. Indonesian language teachers should therefore begin with a precise learning problem, such as weak paragraph coherence or limited revision strategies, rather than begin with the tool and search for a task afterward.

3.3 Lessons from digital storytelling and multimodal composition

Digital storytelling research offers a mature pedagogical comparison because it also involves technology, authorship, and multimodal production. Studies consistently show that digital storytelling is most productive when narrative craft precedes technical editing. Learners benefit from developing plot, perspective, sequence, voice, and audience awareness before combining written language with image, sound, and pacing (Cigerci & Yildirim, 2023; Veyis et al., 2025). Ajabshir (2024) likewise connects digital storytelling with writing development and engagement, while Yu and Wang (2025) emphasize language learning, digital skills, and collaboration.

These findings matter for GenAI because generated media can intensify the temptation to value appearance over meaning. A student can now produce polished images, narration, and scripts rapidly, but a visually convincing story may still contain a weak point of view, stereotypical representation, or borrowed cultural assumptions. Assessment should therefore prioritize narrative coherence, rhetorical purpose, cultural responsibility, and explanation of design choices. Technical quality should remain secondary unless it is itself a stated learning outcome.

Jiang and Hafner (2024) argue that digital multimodal composing requires attention to how modes work together to make meaning. Applied to Indonesian language education, this means students should be able to explain why an image, quotation, soundtrack, layout, or generated element strengthens a particular interpretation. A multimodal text should not become a collage of machine-produced assets. Each element needs an identifiable relationship to the communicative purpose and should be ethically sourced or transparently generated.

The digital-storytelling literature also reinforces the value of local and personal knowledge. Family histories, community practices, oral narratives, regional environments, and contemporary social issues can provide grounding that generic models do not possess. GenAI can be used to generate questions, alternative structures, or comparison texts, but the evidentiary and cultural center of the project should come from human experience, verified sources, and responsible engagement with represented communities. This principle helps prevent cultural flattening and supports the distinctive role of Bahasa Indonesia as a medium for national communication that remains connected to diverse local identities.

3.4 Critical AI literacy, authorship, and academic integrity

The review indicates that critical AI literacy must be taught as part of writing, not as a separate technical orientation. Darwin and Hafner (2022) conceptualize digital literacies as socially situated practices involving participation, power, and identity. Recent GenAI scholarship extends this concern to algorithmic bias, data governance, authorship, and the unequal consequences of automated systems (Kim et al., 2025; Rapanta et al., 2025). Students therefore need more than prompt-writing techniques. They need to understand why

fluent output may be false, whose language patterns are privileged, what information should not be submitted, and how assistance should be disclosed.

Fact checking is the first non-negotiable practice. GenAI can fabricate references, merge details, or present uncertain claims in a confident style. In Indonesian language assignments, students should not cite a source that they have not opened and evaluated. A teacher can require a source-verification table containing the original claim, the verified source, the relevant passage, and the student's decision to accept, revise, or reject the generated suggestion. This moves verification from a vague instruction into visible academic practice.

Authorship is the second practice. Authorship does not require that every word be produced without assistance; writing has long involved dictionaries, peer feedback, templates, and editing tools. The ethical issue is whether the student remains accountable for the argument, evidence, language choices, and final representation. A transparent classroom policy can distinguish permitted uses, such as brainstorming or feedback, from restricted uses, such as generating a final assessed text. Students should disclose the tool, purpose, stage of use, and major changes made after receiving output.

Detection is not a sufficient integrity strategy. Liang et al. (2023) show that AI detectors can disadvantage non-native writers, and generated text can be modified to evade classification. Overreliance on detection shifts attention away from assessment design and may create unjust accusations. More defensible evidence includes developmental writing samples, in-class planning, oral explanation, version history, source notes, and consistent teacher feedback across time. These approaches evaluate learning rather than guessing the origin of a finished paragraph.

Academic-integrity policies also need pedagogical clarity. Cotton et al. (2024) argue that institutions must rethink assessment and communicate expectations in the era of ChatGPT. A blanket prohibition may be difficult to enforce and can prevent students from learning how to use tools responsibly. Unrestricted permission is equally problematic because it can erase the boundary between support and substitution. The review supports a graduated policy in which acceptable uses are specified by task, disclosure is required, and assessment includes process evidence that cannot be outsourced without sustained student participation.

Teacher competence is therefore central. Zhang and Dikilitaş (2025) show the importance of investing in critical digital literacies among novice language teachers, while Kim et al. (2025) emphasize equity and critical ethical literacy. Teachers need opportunities to test tools, compare outputs, develop task-specific prompts, identify risks, and design fair alternatives for students who cannot or do not wish to use GenAI. Without this preparation, classroom rules may be inconsistent, overly punitive, or based on exaggerated assumptions about what the technology can reliably do.

3.5 Equity, access, and the Indonesian language context

GenAI is often described as universally accessible, but meaningful access depends on device quality, connectivity, account restrictions, language support, and the user's capacity to evaluate output. Educational digital transformation research warns that technology can widen existing inequalities when infrastructure and professional support are uneven (Mukul & Büyüközkan, 2023). A responsible Indonesian implementation should therefore provide non-AI pathways that address the same learning outcomes and should not award higher scores simply because some students can access more capable paid systems.

Language quality is another equity issue. Multilingual models may perform unevenly across standard Indonesian, academic registers, regional vocabulary, and mixed-language contexts. Fluent Indonesian output should not be assumed to be culturally accurate or rhetorically appropriate. Teachers can turn this limitation into a comparative literacy task by asking students to identify unnatural diction, unsupported generalization, inaccurate local references, or loss of interpersonal nuance. However, the limitations themselves require systematic empirical testing with Indonesian texts and diverse learner populations.

Privacy must also be treated as a classroom design issue. Students should not upload personal data, confidential school information, unpublished research materials, or identifiable stories about others without consent. Narrative assignments are especially sensitive because they often draw on family, community, or personal experience. An access-sensitive and privacy-aware design allows students to use fictionalized data, teacher-provided excerpts, local offline tools where available, or non-AI peer-review alternatives.

Equity further requires assessment criteria that do not reward standardized machine fluency over authentic development. Students with strong ideas but emerging formal control should not be disadvantaged relative to students who submit highly polished generated prose. Rubrics should allocate substantial weight to evidence quality, coherence of reasoning, authorial voice, revision decisions, and process documentation. Surface correctness remains relevant, but it should be interpreted in relation to the learning goal and the student's demonstrated control.

3.6 The HUMAN-AI process-writing framework

Based on the synthesis, this article proposes a five-stage HUMAN-AI process-writing framework for Indonesian language education. The name emphasizes that human responsibility frames every use of artificial intelligence. The stages are Human grounding, Purposeful prompting, Evidence checking, Authorial reconstruction, and Reflective disclosure. They are iterative rather than strictly linear, but each stage produces an observable learning artifact.

Table 1. The HUMAN-AI process-writing framework for Indonesian language education.

Stage	Primary learning action	Boundary for AI use	Process evidence
Human grounding	Define purpose, audience, genre, context, and initial position.	No generated draft before an original outline or source plan.	Task notes and baseline outline.
Purposeful prompting	Use bounded prompts for questions, alternatives, or rubric-based feedback.	AI supports a specified subtask; it does not produce the final assessed text.	Prompt log and selected output.
Evidence checking	Verify claims, examples, references, and local knowledge independently.	Generated content is never treated as a citable source.	Source-verification notes.
Authorial reconstruction	Rewrite, reorganize, and integrate ideas in the learner's own accountable voice.	Copy-editing cannot replace substantive revision and decision making.	Draft comparison and revision rationale.
Reflective disclosure	Explain the tool, purpose, influence, verification, and limitations.	Use must be transparent and consistent with task rules.	Disclosure statement or oral defense.

Human grounding begins before the tool is opened. Students clarify the task, audience, genre, purpose, prior knowledge, local context, and initial position. They gather or identify credible source material and create a brief outline in their own words. This stage protects the assignment from becoming an exercise in prompt outsourcing and gives the teacher a baseline for later comparison.

The relationship between the HUMAN-AI stages and the reviewed evidence was derived from recurring themes across the corpus. Human grounding reflects findings on learner agency and writing-process pedagogy; purposeful prompting corresponds to studies on guided chatbot interaction and feedback; evidence checking is linked to research on hallucination, source evaluation, and critical AI literacy; authorial reconstruction reflects studies emphasizing revision and ownership; and reflective disclosure follows academic-integrity research concerning transparency and responsible use.

Purposeful prompting means that AI is used for a bounded learning function. Examples include requesting questions that test an argument, asking for two alternative organizational patterns, generating a counterargument for evaluation, or receiving feedback aligned with a rubric. Prompts should include context and constraints, but they should not ask the system to produce the final assessed text. Students keep a concise prompt log so that the role of the tool remains visible.

Evidence checking requires students to verify claims, examples, quotations, and references independently. Generated suggestions are treated as hypotheses, not sources. Students compare output with assigned readings, library materials, credible websites, interviews, or primary observations. Any unsupported statement is removed or rewritten, and every cited source is read directly.

Authorial reconstruction is the stage at which students rewrite, reorganize, and integrate ideas in a voice appropriate to the audience and genre. Copying a generated paragraph and changing a few words is not sufficient. Students should be able to explain the major rhetorical decisions, identify suggestions they rejected, and demonstrate how the final text differs from the generated material. Peer and teacher feedback remain important because they provide situated human interpretation.

Reflective disclosure documents what tool was used, why it was used, what output influenced the text, and how the student verified and transformed that output. A short disclosure statement, revision memo, or oral explanation is usually sufficient. Reflection turns transparency into metacognition: students identify which parts of the writing process they controlled well, where the AI was unreliable, and what they would do differently in a future task.

The framework illustrates one possible assessment orientation. The following weighting is a provisional example requiring validation through future classroom research: communicative purpose and argument, 25%; evidence and accuracy, 25%; organization and authorial voice, 20%; language control and multimodal meaning, 15%; and process documentation, verification, and disclosure, 15%. Teachers can adjust these proportions by level and genre. The essential principle is that technical fluency or visual polish should never conceal weak evidence, unclear authorship, or absent revision.

3.7 Limitations of the review

This review has several limitations. The corpus was limited to 20 peer-reviewed journal articles published between 2022 and 2025, reflecting the rapid development of GenAI research but also restricting the breadth of available evidence. Database coverage, publication language, and indexing practices may have introduced selection bias. The included studies also varied in design, educational level, and research purpose, making direct comparison difficult. In addition, this review did not apply a formal risk-of-bias appraisal because the included literature represented diverse conceptual, qualitative, review, and intervention designs. Future systematic reviews should incorporate broader databases, explicit quality assessment tools, and longitudinal classroom evidence.

3.8 Implications for curriculum, teacher education, and research

At the curriculum level, GenAI should be integrated through explicit learning outcomes rather than occasional tool demonstrations. Students need repeated practice in prompt critique, source verification, comparative revision, disclosure, and discussion of bias. These outcomes can be embedded in existing writing units instead of creating a separate technology subject. For example, an exposition unit can include evaluation of an AI-generated argument, while a narrative unit can include comparison between generic generated description and locally grounded sensory detail.

Teacher education should focus on judgment rather than platform mastery. Specific tools change quickly, but enduring professional capacities include task design, feedback literacy, ethical reasoning, privacy protection, inclusive alternatives, and the ability to distinguish higher-order learning from surface performance. Teachers also need shared institutional guidance so that students are not confronted with contradictory rules across subjects.

The evidence base remains limited in three ways. First, much of the literature concerns English-language learning or higher education. Second, many studies examine perceptions rather than long-term changes in independent writing ability. Third, fast-changing systems make tool-specific findings unstable. Future research in Indonesian language education should therefore use classroom interventions that compare different levels of AI assistance, analyse revision histories, assess delayed independent writing, and include learners from varied regions and infrastructure conditions.

Research should also examine the linguistic quality of generated Indonesian across genres. Relevant questions include whether systems preserve stance and politeness, how they handle standard and regional vocabulary, whether they reproduce cultural stereotypes, and how students negotiate code-mixing. Design-based studies can test the HUMAN-AI framework and refine rubrics for primary, secondary, university, and teacher-education contexts.

4. Conclusion

This systematic literature review examined how recent international research on GenAI, educational chatbots, writing feedback, digital storytelling, multimodal composing, and critical digital literacy can inform Indonesian language education. The synthesis indicates that GenAI is most educationally defensible when it functions as a bounded dialogic scaffold within a visible writing process. It can support planning, questioning, language comparison, and formative feedback, but it should not replace the learner's responsibility for purpose, evidence, organization, voice, and final representation.

The main risks are not limited to cheating. Hallucination, bias, privacy exposure, generic language, cultural flattening, unequal access, and unreliable detection can all undermine fair learning. These risks require explicit critical AI literacy and assessment redesign. Teachers need process evidence, transparent disclosure, source verification, and multiple ways for students to demonstrate competence. AI-detection scores should not be treated as conclusive proof of authorship.

The proposed HUMAN-AI framework translates the review into five classroom stages: human grounding, purposeful prompting, evidence checking, authorial reconstruction, and reflective disclosure. Its central principle is that AI use must remain explainable, verifiable, and subordinate to human communicative purposes. The framework can be adapted to narrative, exposition, argumentation, review, procedure, and multimodal projects. The framework should be understood as a theoretically informed pedagogical model generated from synthesis, not as an empirically validated intervention.

Because direct empirical research in Bahasa Indonesia remains limited, the framework should be tested rather than treated as a proven intervention. Future studies should evaluate independent writing after AI-assisted instruction, compare urban and non-urban implementation, examine linguistic and cultural bias in generated Indonesian, and document how teachers and students negotiate authorship. Responsible adoption will depend less on the novelty of the tool than on the quality of the pedagogy that governs it.

Table 2. Suggested assessment indicators for GenAI-assisted Indonesian writing tasks.

Criterion	Strong evidence	Warning signs
Purpose and argument	Clear claim, audience awareness, logical development, and defensible choices.	Fluent sentences without a stable position or task-specific reasoning.
Evidence and accuracy	Every source is opened, checked, and integrated accurately.	Unverifiable claims, invented citations, or dependence on generated summaries.
Authorial voice	Language and stance are consistent with the learner's demonstrated decisions.	Abrupt style shifts, generic diction, or inability to explain key passages.
Revision process	Drafts show substantive reorganization and reasoned acceptance or rejection of feedback.	Only cosmetic changes or direct copying of generated output.
Disclosure and ethics	Tool, purpose, verification, privacy, and contribution are stated transparently.	Hidden assistance, personal data exposure, or unclear ownership of content.

5. Declarations

Ethical Approval

Ethical approval was not required for this study because it is a systematic literature review analyzing publicly available peer-reviewed journal articles. The research did not involve human participants, animals, or primary data collection requiring ethical clearance.

Data Availability Statement

The data supporting the findings of this study are available within the article and its supplementary materials. Specifically, the list of included studies and their characteristics are presented in Appendix A and Appendix B. Additional information regarding the search strategy can be requested from the corresponding author.

Funding

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

Conflict of Interests

The authors declare that they have no conflict of interests.

Author Contributions

Zathu Restie Utami: Conceptualization, Methodology, Investigation, Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing, Visualization, Supervision, Project Administration. All authors have read and agreed to the published version of the manuscript.

Supplementary Material

No supplementary material is associated with this article beyond the Appendices included in the main text.

Declaration about the Use of Generative AI and AI-Assisted Technologies in the Writing Process

The author declares that no generative AI or AI-assisted technologies were used in the writing of this manuscript.

Acknowledgment

The author would like to thank the editorial team of HANA: Humanities and Academic Narratives in Education for their constructive feedback during the review process. Appreciation is also extended to colleagues who provided insights on the conceptual framework.

References

- Ajabshir, Z. F. (2024). Empowering EFL writing through digital storytelling: A quasi-experimental assessment of CALF measures and multidimensional engagement. *Acta Psychologica*, 252, 104564. <https://doi.org/10.1016/j.actpsy.2024.104564>
- Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20, 43. <https://doi.org/10.1186/s41239-023-00411-8>
- Cigerci, F. M., & Yıldırım, M. (2023). From Freytag pyramid story structure to digital storytelling: Adventures of pre-service teachers as story writers and digital story tellers. *Education and Information Technologies*. <https://doi.org/10.1007/s10639-023-12042-7>
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. <https://doi.org/10.1080/14703297.2023.2190148>
- Darvin, R., & Hafner, C. A. (2022). Digital literacies in TESOL: Mapping out the terrain. *TESOL Quarterly*, 56(3), 865–882. <https://doi.org/10.1002/tesq.3161>
- Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1), 237–257. <https://doi.org/10.1111/jcal.12610>
- Jiang, L. G., & Hafner, C. A. (2024). Digital multimodal composing in L2 classrooms: A research agenda. *Language Teaching*, 58(4), 528–546. <https://doi.org/10.1017/S0261444824000107>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>

- Kim, N., Lee, V. W. Y., Mecham, S., & De Costa, P. I. (2025). Problematizing and addressing equity issues in multiliteracies and the use of generative AI: Why critical ethical literacy matters. *Pedagogies*, 1–14. <https://doi.org/10.1080/1554480X.2025.2557481>
- Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, 54(2), 537–550. <https://doi.org/10.1177/00336882231162868>
- Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 28, 973–1018. <https://doi.org/10.1007/s10639-022-11177-3>
- Liang, W., Yuksekogonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, 11, 9. <https://doi.org/10.1186/s40561-024-00295-9>
- Mogavi, R. H., Deng, C., Kim, J. J., Zhou, P., Kwon, Y. D., Metwally, A. H. S., Tlili, A., Bassanelli, S., Bucchiarone, A., Gujar, S., Nacke, L. E., & Hui, P. (2024). ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. *Computers in Human Behavior: Artificial Humans*, 2(1), 100027. <https://doi.org/10.1016/j.chbah.2023.100027>
- Mukul, E., & Büyüközkan, G. (2023). Digital transformation in education: A systematic review of Education 4.0. *Technological Forecasting and Social Change*, 194, 122664. <https://doi.org/10.1016/j.techfore.2023.122664>
- Rapanta, C., Bhatt, I., Bozkurt, A., Chubb, L. A., Erb, C., Forsler, I., Jandrić, P., Kiryakova, G., Kohnke, L., Lee, D., & others. (2025). Critical GenAI literacy: Postdigital configurations. *Postdigital Science and Education*, 7, 1296–1333. <https://doi.org/10.1007/s42438-025-00573-w>
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10, 15. <https://doi.org/10.1186/s40561-023-00237-x>
- Veyis, F., Gözüyeşil, E., & Ciğerci, F. M. (2025). The effect of digital storytelling on students' language skills performance: A meta-analytic study. *Education and Information Technologies*, 30, 24463–24501. <https://doi.org/10.1007/s10639-025-13722-2>
- Yu, B., & Wang, W. (2025). Using digital storytelling to promote language learning, digital skills and digital collaboration among English pre-service teachers. *System*, 129, 103577. <https://doi.org/10.1016/j.system.2024.103577>
- Zhang, Y., & Dikilitaş, K. (2025). Investing in critical digital literacies involving generative AI: A multiple-case study of second language novice teachers. *International Journal of TESOL Studies*, 250626, 1–25. <https://doi.org/10.58304/ijts.250626>

Appendix A. Review Matrix of Included Studies

Table A1 summarizes the context, focus, and contribution of each of the 20 peer-reviewed journal articles included in the synthesis.

Study	Context / design	Primary focus	Contribution to this review
Ajabshir (2024)	EFL writing; quasi-experimental study	Digital storytelling, CALF measures, and engagement	Shows that technology is strongest when embedded in a staged writing process.
Chan & Hu (2023)	Higher education; student-perception study	Benefits and challenges of generative AI	Supports explicit training, verification, and balanced classroom rules.
Çiğerci & Yıldırım (2023)	Pre-service teachers; digital-storytelling project	Narrative structure and multimodal production	Reinforces narrative craft before technical editing.
Cotton et al. (2024)	Higher education; conceptual analysis	Academic integrity and assessment redesign	Supports process evidence, transparent rules, and authentic assessment.
Darvin & Hafner (2022)	TESOL; conceptual mapping	Critical and socially situated digital literacies	Provides the literacy lens for power, identity, participation, and responsibility.
Huang et al. (2022)	Language learning; systematic review	Chatbot-supported language learning	Shows that outcomes depend on instructional design and learner conditions.
Jiang & Hafner (2024)	L2 classrooms; research agenda	Digital multimodal composing	Highlights purposeful orchestration of words, images, sound, and audience.
Kasneci et al. (2023)	Education; interdisciplinary commentary	LLM opportunities, risks, and mitigation	Establishes the need for fact checking, human oversight, and critical thinking.
Kim et al. (2025)	Multiliteracies; critical conceptual study	Equity and critical ethical literacy in GenAI use	Centers unequal access, bias, and ethical participation.
Kohnke et al. (2023)	Language education; technology review	ChatGPT affordances and required competencies	Identifies interaction, explanation, and ethical digital competence as key.

Appendix B. Review Matrix of Included Studies (continued)

Study	Context / design	Primary focus	Contribution to this review
Kuhail et al. (2023)	Education; systematic review	Design and use of educational chatbots	Confirms that interface novelty alone does not guarantee learning.
Liang et al. (2023)	AI detection; comparative evaluation	Bias against non-native English writers	Warns against using detector scores as conclusive authorship evidence.
Mahapatra (2024)	ESL writing; mixed-method intervention	ChatGPT as formative writing feedback	Provides empirical support for trained, dialogic feedback use.
Mogavi et al. (2024)	Multiple educational sectors; qualitative study	Early-adopter practices and perceptions	Shows simultaneous gains in efficiency and risks of superficial learning.
Mukul & Büyüközkan (2023)	Education 4.0; systematic review	Digital transformation at system level	Connects classroom innovation with infrastructure and institutional readiness.
Rapanta et al. (2025)	Postdigital education; critical synthesis	Critical GenAI literacy	Frames GenAI use as a social, ethical, and pedagogical configuration.
Tlili et al. (2023)	Education; qualitative case study	ChatGPT usefulness, response quality, privacy, and ethics	Documents hallucination, manipulation, privacy, and cheating concerns.
Veyis et al. (2025)	Language skills; meta-analysis	Effects of digital storytelling	Supports multimodal writing as a language-learning process.
Yu & Wang (2025)	Pre-service teachers; empirical study	Digital storytelling, language, skills, and collaboration	Links writing development with digital collaboration and reflection.
Zhang & Dikilitaş (2025)	Novice language teachers; multiple case study	Critical digital literacies involving GenAI	Shows that teacher preparation is essential for responsible implementation.