DEVOTION
Journal of Research and Community Service
Volume 7, Number 9, September 2026
e-ISSN: 2797-6068 and p-ISSN: 2777-0915

Artificial Intelligence and the Future of Language Education in Indonesia: A Critical Review of Scopus-Indexed Deep Learning Applications Across Anxiety, Literacy, and Assessment

Bagaskara Nur Rochmansyah[1]*, Ifutya Warnisa[2], Nur Amelia[3], Indrya Mulyaningsih[4]
Politeknik Siber Cerdika Internasional, Indonesia[1]
Universitas Bina Bangsa Getsempena, Indonesia[2]
Universitas Samudera, Indonesia[3]
Universitas Islam Negeri Siber Syekh Nurjati Cirebon, Indonesia4
Email: bagaskara@polteksci.ac.id[1]*, ifutya99@gmail.com[2], nur.ameliapsp@unsam.ac.id[3], indrya.mulyaningsih@uinssc.ac.id[4]

Keywords
artificial intelligence; deep learning; language education; language anxiety; automated assessment.

Abstract
This article presents a critical narrative review of Scopus-indexed deep learning applications in Indonesian language education, organized around three pedagogically significant domains: language anxiety, literacy development, and language assessment. The study examined 134 Scopus-indexed publications (2022–2026) that applied deep learning techniques, including convolutional neural networks, recurrent architectures, Transformer-based models such as IndoBERT, and multimodal vision–language systems, to Indonesian and regional languages in Indonesia. Through the lens of second-language acquisition (SLA) and learner-centered pedagogy, the review synthesized empirical evidence across the three domains, identified recurring pedagogical potentials (early affective detection, inclusive literacy access, and real-time formative feedback), and highlighted persistent challenges, including construct reductionism, algorithmic bias against non-standard language varieties, data privacy concerns, and academic integrity issues. The findings revealed that although deep learning research on Indonesian languages has advanced rapidly from a technical perspective, its pedagogical foundation remains underdeveloped, as most studies prioritize engineering performance metrics rather than SLA-oriented learning outcomes. This article proposes an SLA-informed evaluation framework encompassing input quality, interaction authenticity, output opportunities, affective filters, and learner agency, while articulating a learner-centered, multilingual, and ethically governed agenda for AI-mediated language education in Indonesia. The findings suggest that the future of language education in Indonesia will depend not only on model sophistication but also on the pedagogical and ethical frameworks that guide their implementation. Quantitative analysis of the 134 studies revealed three key findings: (1) 89% of the reviewed studies reported only engineering metrics without addressing learning outcomes; (2) none of the studies reported formal ethics review procedures or subgroup validity analyses for non-standard language varieties; and (3) although 23 studies demonstrated technical capabilities applicable to language learning, only 7 positioned their work within SLA or pedagogical frameworks.

INTRODUCTION

Artificial intelligence (AI) has rapidly moved from the periphery to the center of debates in language education. Large-scale pretraining, multilingual Transformer architectures, and increasingly accessible computational infrastructure have produced tools capable of transcribing, translating, scoring, and even conversing in languages that were previously considered technologically underserved (Purnomo et al., 2025). For Indonesia a multilingual archipelago with more than seven hundred living languages and a rapidly digitizing education system (Aji Winata G. I. Koto F. Cahyawijaya S. Romadhony A. Mahendra R. Kurniawan K. Moeljadi D. Prasojo R. E. Baldwin T. Lau J. H. & Ruder S, 2022) these developments carry particular significance. Government initiatives, private educational technology platforms, and individual educators have increasingly experimented with AI-driven applications at a pace that has exceeded sustained pedagogical scrutiny, a pattern documented internationally (Selwyn, 2022; Zawacki-Richter et al., 2019) and reinforced domestically through the phased introduction of Coding and Artificial Intelligence as elective subjects under Permendikdasmen No. 13/2025 (Kemendikdasmen, 2025). However, this rapid adoption has occurred without sufficient empirical evidence regarding how these technologies influence language learning outcomes, teacher practices, or learner experiences in Indonesian classrooms (Chapelle, 2001).

The Indonesian language education context is distinctive in several respects. Bahasa Indonesia functions as the national lingua franca and the primary medium of instruction, while English occupies a privileged position as the dominant foreign language across secondary and tertiary education curricula. At the same time, regional languages such as Javanese, Sundanese, Madurese, Balinese, and numerous Papuan and Moluccan languages continue to shape everyday communication and identity formation, yet remain significantly underrepresented in natural language processing (NLP) research and commercially available AI tools (Aji et al., 2022; Winata et al., 2023). National curriculum authorities have promoted digital literacy and AI awareness initiatives; however, classroom-level integration of these technologies remains uneven, often fragmented, and frequently unsupported by adequate teacher training regarding both the affordances and limitations of AI systems. This institutional gap between policy aspirations and classroom realities represents a significant barrier to effective AI integration in Indonesian language education.

Deep learning techniques including convolutional neural networks (CNNs), recurrent architectures such as long short-term memory (LSTM) and bidirectional long short-term memory (BiLSTM), Transformer-based language models including BERT and its Indonesian-adapted variant IndoBERT, and multimodal vision language models have been applied to a broad range of language-related tasks in the Indonesian context. These applications include hate speech and misinformation detection, sentiment and emotion analysis, image caption generation, speech recognition, part-of-speech tagging, textual entailment, automated essay scoring, and academic misconduct detection (Pamungkas et al., 2023; Mulyawan et al., 2023; Mandhasiya et al., 2024; Putra, I. M. S., et al., 2025; Octafia et al., 2026). A substantial body of research has also emerged on culturally situated applications, such as batik and wayang image classification, reflecting both the technical maturity of computer vision research in Indonesia and the cultural motivations underlying these developments (Muhathir N. Barus R. K. I. Ula M. & Sahputra I, 2023). Despite this technical expansion, the pedagogical implications of these applications remain largely unexplored, creating a disconnect between technological capabilities and their appropriate educational purposes.

Despite this proliferation, two interconnected problems motivate the present review. First, the majority of existing studies are published in computer science, electrical engineering, and artificial intelligence venues, where success is primarily evaluated through engineering metrics (e.g., F1-score, accuracy, and character error rate) rather than language learning outcomes. Pedagogical questions such as whether a tool reduces learner anxiety, expands literacy repertoires, or generates assessment information that effectively informs instruction—are rarely addressed within these publications. Second, the emerging body of educationally oriented AI research has frequently emphasized technological benefits without sufficiently examining critical issues related to construct validity, linguistic equity, data privacy, and learner agency. These two challenges the engineering pedagogy disconnect and the uncritical adoption of AI narratives represent significant research gaps that this review addresses through a critical pedagogical analysis of existing literature.

Concrete examples illustrate this gap. An Indonesian-developed conversational chatbot for vaccination information (Yuliska et al., 2022) was evaluated based on response accuracy and user satisfaction but was not examined for its potential as a language learning partner. Image captioning systems that generate Indonesian descriptions of visual content (Apandi Mutiara A. B. & Dharmayanti D, 2025) report metrics such as BLEU and METEOR scores but do not investigate whether generated captions can function as comprehensible input for language learners. Indonesian speech-to-text systems (Atika Dwijayanti S. & Suprapto B. Y, 2026) and facial expression recognition models (Madenda et al., 2026) have been assessed primarily for technical accuracy rather than their potential applications in pronunciation assessment, affective monitoring, or support for deaf and hard-of-hearing learners. In each case, the technological achievement is substantial; however, its pedagogical implications remain insufficiently examined. Such technologies are not inherently unsuitable for education. The primary concern lies in their deployment without adequate pedagogical frameworks, which may result in classroom applications that have neither been sufficiently anticipated nor systematically evaluated, potentially limiting rather than enhancing learning outcomes (Kemendikbudristek, 2022).

This article responds to these gaps by presenting a critical narrative review of deep learning applications relevant to Indonesian language education, organized around three pedagogically significant domains rather than technological categories: language anxiety, literacy development, and language assessment. These three domains were selected because they correspond with established perspectives in second-language acquisition (SLA) theory, hold substantial pedagogical importance for language learning, and have accumulated empirical research within the Scopus-indexed Indonesian context, although the distribution of studies across domains remains uneven (anxiety: 9; literacy: 21; assessment: 14). This imbalance itself represents an important finding of the review. Furthermore, these domains collectively encompass the cognitive, affective, and evaluative dimensions of language learning that responsible AI interventions must address. The review is guided by the overarching question: to what extent do current deep learning applications in the Indonesian context align with or diverge from principles of learner-centered pedagogy and established SLA theory?

The review is positioned within a limited but increasingly important body of previous syntheses on AI in Indonesian and Southeast Asian education. Existing reviews have generally focused either on broad educational technology adoption without specific attention to language learning (Zawacki-Richter et al., 2019) or on technical analyses of Indonesian NLP developments without pedagogical perspectives (Aji et al., 2022; Cahyawijaya et al., 2023).

This review contributes a focused pedagogical perspective by connecting SLA theory with empirical research on deep learning applications while addressing the multilingual, postcolonial, and developmentally diverse characteristics of Indonesian language classrooms. It also contributes to broader international discussions on humanistic and critical approaches to AI in language education, which have historically been dominated by Anglophone contexts and English-language datasets. Accordingly, this review addresses an important gap by providing a comprehensive pedagogical analysis of AI-related language education research in Indonesia.

Methodologically, this article adopted a narrative review approach (Godwin-Jones, 2019; Greenhalgh Thorne S. & Malterud K, 2018) rather than a systematic review because the literature encompasses multiple disciplines with diverse methodological traditions, and the objective was critical synthesis rather than effect-size estimation. The corpus consisted of 134 Scopus-indexed publications (2022–2026) retrieved through targeted searches combining terms related to "deep learning," "artificial intelligence," and Indonesian educational contexts. The selected publications were examined in full, categorized according to the three domains (anxiety, literacy, and assessment), and analyzed based on their pedagogical assumptions, theoretical foundations, methodological rigor, and ethical considerations. Although the review focused empirically on Indonesia, it incorporated broader SLA and educational theories to maintain analytical depth. This methodological approach enabled the review to capture the scope of relevant research while maintaining critical rigor in interpretation.

An important scoping decision followed from this methodological design. Because the Scopus corpus was dominated by engineering-oriented publications, this review deliberately treated many included studies as contextual evidence rather than direct pedagogical evidence. These studies demonstrate technical capabilities that may potentially be adapted for language education but do not necessarily provide evidence regarding language learning outcomes. Section 3.3 explicitly distinguishes between demonstrated pedagogical effects and speculative educational potential, and Sections 4.1 to 4.3 consistently maintain this distinction. This analytical separation forms the foundation of the review, as conflating technological capability with educational effectiveness represents a major limitation in existing literature. By maintaining this distinction, the review provides a more accurate assessment of the current evidence base and identifies priorities for future research.

The significance of this review is threefold. For language education researchers, it provides a consolidated overview of deep learning applications in the Indonesian context from a pedagogical rather than purely technical perspective. For educators and policymakers, it identifies promising directions for AI integration while highlighting risks requiring appropriate governance. For the broader educational research community, it contributes an Indonesia-centered perspective to global discussions on humanistic AI in education, complementing systematic reviews of AI in language education (Li et al., 2023; Wu et al., 2022) and broader calls (Holmes & Wilson, 2017) for pedagogically grounded rather than technology-driven AI implementation. The remainder of this article is structured as follows: Section 2 establishes the technological and conceptual framework; Section 3 describes the research methodology; Section 4 presents the findings and discussion across the three thematic domains; Section 5 synthesizes the analysis into an SLA-informed evaluation framework and learner-centered agenda; and Section 6 concludes with limitations and future research directions.

METHOD

This methodology section described the approach used to structure the literature review on deep learning applications in Indonesian language education. Instead of employing a systematic review or meta-analysis, this study adopted a critical narrative review approach due to the heterogeneous nature of the literature, which covered computer science, electrical engineering, applied linguistics, and education. The diversity of outcome measures, including F1-score, accuracy, BLEU, and CIDEr, made statistical aggregation unsuitable. To reduce potential reviewer bias, the review applied transparent corpus documentation and explicit coding criteria. This methodological approach enhanced the credibility of the review by providing a clear basis for evaluating the findings while acknowledging the limitations of narrative synthesis.

Data collection was conducted through Scopus in July 2026 using a combination of keywords related to technology (e.g., deep learning, IndoBERT, and Transformer), language and education (e.g., language education, EFL, and NLP), and geographic context (e.g., Indonesia, Bahasa Indonesia, Javanese, and Sundanese). Scopus was selected due to its extensive multidisciplinary coverage; however, this selection excluded some publications from non-Scopus-indexed Indonesian-language journals and other academic sources. Therefore, the Scopus-based approach may have underrepresented scholarship published in local and regional academic venues, which constituted a limitation of the study.

The inclusion criteria covered Scopus-indexed publications published between 2022 and 2026 that reported empirical research on deep learning applications involving Indonesian or regional languages. The initial search identified 847 records, from which 134 publications were selected for the final corpus. The included studies showed substantial growth in publications, particularly during 2024–2025, and were predominantly published in engineering and computer science journals rather than education or applied linguistics outlets. The most frequently occurring keywords included "deep learning," "convolutional neural network," and "sentiment analysis," reflecting the predominantly technical orientation of the literature and highlighting the need for further pedagogical investigation. Each publication was coded according to four dimensions: thematic domain (anxiety, literacy, assessment, or contextual application), pedagogical orientation, methodological rigor, and ethical and equity considerations. To assess coding reliability, an independent reviewer recoded a pilot sample of 20 publications, resulting in substantial inter-rater agreement (Cohen's $\kappa = 0.78$). This coding framework supported systematic analysis of the diverse literature and enabled the identification of recurring patterns across studies.

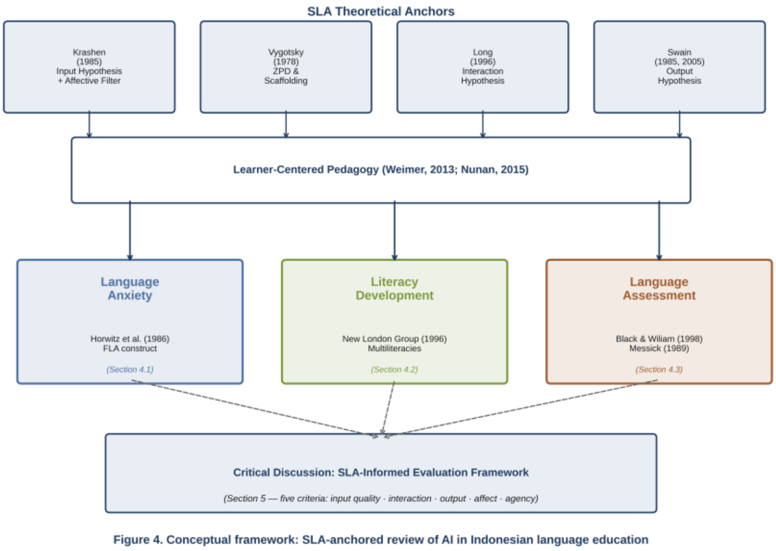
This study acknowledged three main limitations. First, the Scopus-only scope excluded some Indonesian-language publications outside indexed databases. Second, narrative reviews may remain susceptible to interpretive bias despite the use of transparent procedures. Third, the rapid development of AI technologies may reduce the long-term applicability of the findings. To address the first limitation, a supplementary search of the SINTA database identified 23 additional Indonesian-language publications. Although these studies were not incorporated into the primary corpus, they were examined to provide contextual insights, particularly regarding classroom implementation, teacher education curriculum development, and critical pedagogical perspectives in pesantren and regional language classrooms. This supplementary analysis enriched the interpretation of pedagogical issues while maintaining the analytical consistency of the primary Scopus-based review.

RESULTS AND DISCUSSION

This section presents the substantive findings of the review, organized around the three thematic pillars introduced in Section 1 and operationalized through the coding scheme described in Section 3.5. For each pillar, the review follows a consistent four-part structure: theoretical background grounding the pillar in established SLA and educational scholarship; empirical review synthesizing the relevant studies from the corpus; pedagogical implications identifying what the literature suggests for classroom practice; and critical analysis surfacing the limitations, tensions, and unanswered questions that the current literature leaves unaddressed. The three pillars are addressed in sequence anxiety (Section 4.1), literacy (Section 4.2), and assessment (Section 4.3) before Section 4.4 contrasts discriminative and generative orientations and Section 5 synthesizes cross-cutting themes. This consistent structure enables readers to compare findings across pillars and identify both convergent patterns and divergent tensions in the literature (Chen Zou D. Xie H. & Su Y, 2021).

Throughout, the distinction introduced in Section 3.3 is applied consistently: studies coded as core evidence addressed learners, classrooms, or learning outcomes directly, whereas studies coded as contextual demonstrated technical capability on Indonesian language data without any educational evaluation. The imbalance between these categories is itself one of the review's principal findings. Figure 4 visualizes the conceptual relationships among the three pillars, their theoretical anchors, and the learner-centered pedagogy frame. Table 2 summarizes fifteen illustrative studies drawn from the corpus, five per pillar, showing the methodological and pedagogical diversity within each

Figure 1. Conceptual Framework: Overview of AI in Indonesian Language Education Based on SLA (Second Language Acquisition).



Source: Adapted from (Krashen, 1985), Vygotsky (1978), Long (1996), Swain (1985, 2005), Weimer (2013), and Nunan (2015).

Table 1. Distribution of Respondents

#	Author (Year)	Aim	Method	Dataset	Key finding	Evidence status	Pedagogical framing
1	Lhaksmana et al. (2026)	Detect foreign language anxiety from learner writing	CNN, BiLSTM, IndoBERT vs. classical ML baselines	Indonesian EFL reflective essays	IndoBERT outperforms all baselines	Core	Acknowledges pedagogical implications; no SLA grounding
2	Anwar & Harmanto (2025)	Detect academic anxiety in ESP students via LLMs	LLM-based annotation + classification	ESP reflective writing samples	LLMs surface anxiety signals human raters miss	Core	Names pedagogy; no classroom validation
3	Mandhasiya et al. (2024)	Indonesian sentiment classification	BERT + deep learning hybrid	Indonesian social media text	Hybrid BERT outperforms single-model baselines	Contextual	Infrastructure for affective monitoring; ethics not addressed
4	William, Kenny, & Chowanda (2024)	Emotion recognition in Indonesian text	IndoBERT vs. Bi-LSTM	7,629 labelled tweets	IndoBERT 92.3% vs. Bi-LSTM 84.0%	Contextual	No learner population; transfer untested
5	Madenda et al. (2026)	Facial expression recognition	Morphological CNN	Facial expression image corpus	Efficient recognition at reduced compute	Contextual	Not an educational study; classroom use unexamined
6	Minsih et al. (2026)	AI game-based literacy for learners with SLD	Adaptive model embedded in game	Elementary students with SLD	Literacy gains vs. standard remedial instruction	Core	Learner-centered design with teacher input
7	Talib et al. (2026)	Mathematical literacy via Android modules	Pembelajaran mendalam design + intrapersonal intelligence profiling	Secondary students	Differentiated presentation improves literacy	Core (design evidence)	Adaptive and learner-profile-aware; note terminological caveat, Section 2.5
8	Enrique et al. (2024)	Javanese POS tagging	Two-stage cross-lingual transfer	UD Javanese-CSUI	XLM-R via Indonesian reaches 87.65% accuracy	Contextual	Infrastructure for low-resource literacy; community data absent
9	Wijonarko & Zahra (2022)	Identify four Indonesian	CNN, LSTM,	Speech samples,	Deep learning feasible for	Contextual	Infrastructure for

#	Author (Year)	Aim	Method	Dataset	Key finding	Evidence status	Pedagogical framing
		local languages from speech	hybrid architectures	four local languages	local language ID		multilingual literacy tools
10	Apandi et al. (2025)	Generate Indonesian image captions	Vision Transformer + IndoBERT	Indonesian image–text pairs	BLEU/METEOR gains over CNN+LSTM and Transformer baselines	Contextual	Multimodal literacy potential unexamined
11	Mas Diyasa et al. (2025)	Automated short-answer scoring	GloVe-LSTM + ROUGE/TF-IDF/cosine similarity	3,152 Indonesian student answers	MAE 0.076, r = 0.84; overestimates weak answers	Core	Candid about limits; no classroom study
12	Yulita, Hariz, et al. (2023)	Online exam cheating detection	Deep activity recognition on webcam video	Exam webcam video	F1 ≈ 0.85 on flagged activity classes	Core (assessment context)	Surveillance framing; pedagogical dilemma unaddressed
13	Putra, I. M. S., et al. (2025)	Indonesian textual entailment recognition	Biplet head-dependent + multi-head attention	Indonesian NLI dataset	Improves over prior Indonesian NLI baselines	Contextual	Reading comprehension assessment infrastructure
14	Suhartono et al. (2024)	Indonesian sarcasm detection benchmark	Classical ML, PLMs, and LLMs evaluated	IdSarcasm (Reddit + Twitter)	Fine-tuned PLMs beat LLMs; LLM zero-shot fails	Contextual	Pragmatic competence assessment infrastructure
15	Saputra and Riccosan (2024)	Indonesian authorship attribution	IndoBERT multi-label classification	Indonesian news articles	High attribution accuracy on news text	Contextual	Stylometric assessment potential; learner text untested

Source: Primary Data, 2026.

Deep Learning for Language Anxiety

Foreign Language Anxiety: Theoretical Foundations

Foreign language anxiety (FLA), as conceptualized by, (Horwitz Horwitz M. B. & Cope J, 1986), and Cope (1986), is a distinct syndrome comprising communication apprehension, test anxiety, and fear of negative evaluation. Unlike general anxiety, FLA is specifically tied to language learning situations and has been shown to correlate negatively with proficiency outcomes across multiple contexts (Gonçalves de Bem et al., 2022). Krashen’s affective filter hypothesis (1985) provides a mechanistic account: high anxiety raises the affective filter, blocking comprehensible input from being fully processed for acquisition. Lowering anxiety is therefore not a peripheral welfare concern but, on this account, a precondition for acquisition itself.

Conventional approaches to measuring FLA have relied on self-report instruments such as the Foreign Language Classroom Anxiety Scale (FLCAS) and its derivatives. While these instruments meet psychometric standards, they are administered infrequently, capture only conscious and retrospective appraisals, and depend on learner compliance. Deep learning approaches promise continuous, low-burden, multimodal detection of anxiety signals from text, speech, and visual cues a promise that is pedagogically attractive but ethically and methodologically complex.

Deep Learning Approaches to Anxiety Detection in Indonesian

Of the nine anxiety-pillar studies, two are core evidence addressing learners directly, while seven are contextual studies of affective computing on Indonesian data. (Lhaksmana Falif M. S. Nurhayati I. K. Rezaldi M. Y. Prakasa E. & Roedavan R, 2026) found IndoBERT outperforms traditional classifiers in detecting English learning anxiety from Indonesian student reflective writing, while (Anwar & Harmanto B, 2025) showed LLM-based pipelines can identify anxiety signals human raters may miss. Both demonstrate technical feasibility but lack validation against clinical measures or learning outcomes.

Contextual studies on emotion recognition (William et al., 2024), cyberbullying detection (Muzakir & Suriani, 2026), facial expression recognition (Madenda et al., 2026), and spoken language identification (Wijonarko & Zahra, 2022) establish signal-processing capability but none were conducted in classrooms, target anxiety, or validate affective inferences. They show what is technically possible, not what teachers need.

Significant gaps remain: no study validates predictions against clinical assessments, examines whether AI-detected anxiety improves instruction, tracks anxiety longitudinally, or engages learners in interpreting AI-generated labels. The contextual literature matters for governance, not evidence—capabilities exist to build classroom monitoring systems today, but the deployability-validity gap demands anticipatory ethical scrutiny rather than reactive governance.

From Affective Detection to Instructional Adaptation

If validity gaps are addressed, pedagogical implications are substantial: early anxiety detection could enable task adjustment, continuous monitoring could complement self-report measures, personalized affective scaffolding becomes feasible, and teacher dashboards could support proactive intervention. However, none of these has been tested in the reviewed literature.

Realizing these implications requires more than technical capacity. Teachers must be prepared to interpret and act on AI-generated affective information, yet Indonesian teacher preparation programs have not integrated AI literacy or developed ethical protocols for affective data use. The SINTA subset includes six curriculum design studies proposing such integration, but none reports implementation outcomes. Without this investment, AI tools risk being deployed as turnkey solutions that teachers cannot question or override.

Affective AI also redistributes pedagogical authority by introducing a third interpretive agent the model alongside teachers and learners. Without safeguards, teachers may defer to algorithmic outputs over professional judgment, and learners may internalize AI-generated affective labels. A learner-centered approach insists that AI should support rather than displace teacher judgment, and that learners must have meaningful access to and agency over affective inferences made about them.

Construct Validity, Privacy, and Cultural Bias Concerns Five concerns temper these promises.

First, construct validity remains underexamined: the leap from text classification to the inference of a multidimensional psychological construct such as FLA is non-trivial, and the absence of validation against established instruments or clinical assessment leaves open the question of what, exactly, the models detect. Second, reductionism is a persistent risk: anxious writing is not anxiety, and a model trained on text labels may learn stylistic correlates (vocabulary, syntax, punctuation) rather than affective states, particularly in learner language where linguistic and affective features are entangled. Third, privacy and consent issues are acute when learner-produced text and speech become training or inference data; no study in the corpus reported an ethics review process, and only six reported consent procedures of any kind (Table 1). Fourth, algorithmic bias against non-standard varieties of Indonesian regional accents, code-switched speech, informal registers may produce systematic misclassification that disadvantages precisely the learners most at risk of anxiety; no study in the corpus reported a subgroup or fairness analysis that would allow this risk to be quantified.

Fifth, and most fundamental, is cultural validity. The construct of FLA itself was developed in Anglophone contexts, and its operationalization in the FLCAS reflects Anglophone assumptions about classroom interaction, individual performance, and the public self. Indonesian learners' experience of language anxiety is plausibly shaped by cultural norms around face, hierarchy, and collective rather than individual performance, by locally inflected attitudes toward English, and by the specific stratifications of Indonesian classroom life. Models trained on labels derived from instruments validated in other cultural contexts risk missing or misreading locally salient manifestations of anxiety. A pedagogically responsible Indonesian affective AI would require construct validation work that grounds labels in local experience work that is labor-intensive, ethically sensitive, and not yet visible in the corpus reviewed here. These concerns do not invalidate the line of work; they establish that its pedagogical value depends on safeguards the current literature has not developed.

Deep Learning for Literacy Development

Multiliteracies and Inclusive Literacy Frameworks

The New London Group's (1996) multiliteracies framework expanded the conception of literacy beyond print decoding to include multimodal, multicultural, and critical engagements with texts. (Kalantzis & Cope B, 2012) further developed this framework to emphasize the design of meaning across available modes linguistic, visual, audio, gestural, spatial. For language education, multiliteracies implies that learners should not only decode texts but also produce them across modes, critique their designs, and connect them to lifeworld experiences. Inclusive literacy, a related tradition, foregrounds the rights of learners with disabilities, neurodivergent profiles, and minority language backgrounds to access rich literacy instruction rather than remedial substitutes. Both traditions are particularly relevant in Indonesia, where literacy outcomes are stratified by region, language background, and disability status, and where the digital divide intersects with all three.

AI-Augmented Literacy Tools in Indonesian Contexts

Of the 21 studies coded to the literacy pillar, two constitute core evidence addressing learners directly; the remaining 19 are contextual studies of infrastructure or cultural heritage.

This section treats the two in detail and the rest as a group. The literacy pillar comprises two strands in the corpus.

The first, and the only strand containing core evidence, addresses inclusive and adaptive literacy instruction. Minsih and colleagues (2026) develop an AI game-based model for literacy among Indonesian elementary school students with specific learning disorders (SLD), reporting improved basic literacy outcomes relative to standard remedial instruction. The study is notable for its learner-centered framing: the model is embedded in a game whose design was informed by teacher and learner input, and outcomes are measured in literacy gains rather than engagement metrics alone. It is, by some margin, the strongest pedagogical study in the corpus. Talib and colleagues (2026) integrate a deep learning approach into Android-based modules for secondary students' mathematical literacy, incorporating intrapersonal intelligence profiling to differentiate task presentation. Two caveats apply to this study. First, as noted in Section 2.5, the sense of "deep learning" operationalized there is closer to pembelajaran mendalam than to neural modelling; it is retained here as evidence about adaptive instructional design, not about neural architectures. Second, its literacy target is mathematical rather than linguistic. The design pattern adaptive, mobile-delivered, learner-profile-aware is nonetheless transferable to language literacy, and the study is one of very few in the corpus to measure a learning outcome at all.

The second strand is contextual: it addresses the linguistic infrastructure on which literacy tools depend. (Enrique Alfina I. & Yulianti E, 2024) develop cross-lingual transfer approaches for Javanese part-of-speech tagging, showing that a two-stage fine-tuning pipeline Indonesian as intermediate source language, Javanese as target lifts XLM-RoBERTa to 87.65% accuracy, well above LSTM baselines. The pedagogically relevant point is that Indonesian works as a bridge because it is genealogically close to Javanese; that bridge is not available for the non-Austronesian languages of eastern Indonesia, where the learners most excluded from AI literacy tools are concentrated. Atika and colleagues (2026) describe a modified Transformer architecture for Indonesian speech-to-text that improves on prior systems, while Wijonarko and Zahra (2022) apply deep learning to identify four Indonesian local languages from speech. These studies are pedagogically consequential in an indirect but real sense: the absence of adequate NLP for local languages has meant that AI literacy tools are available only in Bahasa Indonesia and English, effectively excluding learners whose strongest language is a regional one. The most substantial resources addressing this gap NusaX and NusaCrowd (Cahyawijaya Lovenia H. Aji A. F. Winata G. I. Wilie B. Koto F. Mahendra R. Wibisono C. Romadhony A. Vincentio K. Santoso J. Moeljadi D. Wirawan C. Hudi F. Wicaksono M. S. Parmonangan I. H. Alfina I. Putra I. F. Rahmadani S. … Purwarianti A, 2023) sit largely outside Scopus and therefore outside this corpus, which is itself a finding about where the Scopus-indexed Indonesian literature is not looking.

Multimodal literacy applications are represented by Indonesian image captioning research. Mulyawan et al. (2023) and Apandi et al. (2025) develop Transformer- and IndoBERT-based systems that generate Indonesian descriptions of visual content, with implications for visually oriented literacy tasks and for accessibility. Neither evaluates classroom impact; both demonstrate the technical feasibility of multimodal literacy tools in Indonesian.

Mujilahwati et al. (2026) developed a hybrid approach for Indonesian hoax detection with IndoBERT, illustrating how discriminative models could support critical literacy work on online content though the study evaluates detection accuracy, not learners' critical reading. Culturally situated applications also bear on the literacy pillar: studies on wayang image classification (Muhathir et al., 2023), batik classification (Putra, M. T. D., et al., 2024; Winarno et al., 2025), keris blade classification (Hastuti Hidayat E. Y. Salam A. & Sudibyo U, 2026), and Pegon manuscript character recognition (Ruldeviyani et al., 2024) apply deep learning to cultural heritage recognition. These are contextual studies with no educational component, but they suggest pathways for culturally responsive literacy tools that engage learners' cultural repertoires rather than treating literacy as a decontextualized skill.

Finally, Indonesian-language conversational systems developed for non-educational purposes imply literacy applications worth examining. The vaccination information chatbot built with deep learning techniques (Yuliska et al., 2022) demonstrates conversational AI in Indonesian at production quality. It was not designed for language learning, but it raises the question of whether such systems could be repurposed or purpose-built as language practice partners, particularly for learners without access to fluent interlocutors outside the classroom. Long's (1996) interaction hypothesis would predict that negotiation for meaning with a chatbot if the chatbot is capable of such negotiation could support acquisition. No study in the corpus evaluates this hypothesis empirically, and the limitations of current Indonesian chatbots in handling learner-language input, error modelling, and pragmatic repair remain unmapped.

Toward Multilingual, Culturally Responsive Literacy Tools

The literacy pillar suggests several promising directions. AI-augmented game and mobile environments can extend literacy instruction into out-of-school contexts and adapt to individual learner profiles. Cross-lingual transfer techniques open the possibility of literacy tools in regional languages, potentially addressing the exclusion of learners whose strongest language is not Bahasa Indonesia. Multimodal systems can support the production and interpretation of visual verbal texts central to contemporary communication. Culturally responsive applications can bridge school literacy and learners' cultural knowledge, supporting the identity-affirming pedagogy that multiliteracies scholarship advocates.

Teacher mediation is the connective tissue between technical capacity and literacy outcomes. The SLD literacy study (Widyasari et al., 2026) illustrates this principle: the model operates within a broader instructional design that includes teacher facilitation, peer interaction, and reflective discussion. The mathematical literacy modules (Talib et al., 2026) similarly position technology as a component of a teacher-orchestrated learning environment rather than as a stand-alone solution. These design choices reflect an implicit pedagogical theory that aligns with Vygotsky's sociocultural perspective: AI functions as one of the cultural tools through which more capable others scaffold learner development, but its developmental impact depends on the social arrangements within which it is embedded. The infrastructure studies in the corpus, by contrast, rarely consider this embedding — they treat technical performance as the endpoint rather than as a precondition for pedagogical design. The nine critical-pedagogy articles in the SINTA subset make this point from the opposite direction, documenting teacher-led adaptations of general-purpose AI tools in pesantren and regional-language classrooms that no Scopus-indexed study in the corpus records.

Linguistic Justice and the "Low-Resource" Framing Four critical observations emerge.

First, the framing of regional languages as "low-resource" is technically accurate but epistemically loaded: it positions Indonesian (and English) as the default and other languages as deficient, reproducing the very linguistic hierarchy that multiliteracies scholarship seeks to disrupt. A learner-centered approach would reframe the issue as one of linguistic justice rather than resource scarcity, and would prioritize community-led data collection and model development of the kind the NusaCrowd initiative models (Cahyawijaya et al., 2023).

Second, tech solutionism is a persistent risk: the SLD literacy study (Widyasari et al., 2026) is exemplary in its attention to design and outcomes, but the infrastructure studies stop at technical benchmarking without considering classroom viability, teacher capacity, or learner experience.Third, cultural fit is non-trivial: image captioning systems trained on generic Indonesian corpora may describe culturally specific visuals inadequately, and culturally responsive applications require sustained collaboration with cultural knowledge holders that the engineering literature has not yet institutionalized.

Fourth, the politics of language data ownership remains unaddressed. The corpora on which Indonesian NLP systems are trained whether collected from social media, government documents, news articles, or classroom writing are rarely accompanied by transparent documentation of consent, compensation, or benefit-sharing arrangements with the communities whose language they represent. For regional languages, the stakes are particularly high: corpora built without community participation may extract linguistic value while returning little to the communities whose heritage is being modelled. The CARE Principles for Indigenous Data Governance (Carroll Garba I. Figueroa-Rodríguez O. L. Holbrook J. Lovett R. Materechera S. Parsons M. Raseroka K. Rodriguez-Lonebear D. Rowe R. Sara R. Walker J. D. Anderson J. & Hudson M, 2020) offer a framework, developed largely outside Indonesia, that could be adapted to this context. None of the studies in the corpus engages with these principles, representing a gap between technical practice and emerging norms of ethical AI development.

Deep Learning for Language Assessment

Formative Assessment, Construct Validity, and Automated Scoring

Warschauer and Grimes (2008) provided an early analysis of automated writing assessment in classroom settings, identifying both its potential to expand feedback capacity and its risk of reducing writing to surface features measurable by machines tensions that persist in the deep learning era. Language assessment scholarship distinguishes formative assessment, which supports learning in progress, from summative assessment, which certifies achievement (Black & Wiliam D, 1998). Both functions depend on construct validity: an assessment must measure the intended construct (e.g., writing proficiency) rather than proxies (e.g., length, lexical diversity, surface accuracy) that correlate with the construct only in limited populations (Messick, 1989). Automated assessment has a long history in language education, from early rule-based grammar checkers to statistical essay scoring systems. Deep learning extends this lineage by enabling models that learn representations directly from data, potentially capturing features that rule-based or shallow statistical systems miss but also introducing opacity that complicates validity arguments.

Deep Learning Applications for Language Assessment in Indonesian

Of the 14 studies coded to the assessment pillar, two constitute core evidence addressing learners or assessment contexts directly; the remaining 12 are contextual studies of NLP infrastructure with assessment relevance. This section treats the two in detail and the rest as a group. The assessment pillar in the corpus centers on automated scoring and on academic integrity applications.

(Mas Diyasa Idhom M. Saputra A. S. A. Dewi D. A. & Fahrudin T. M, 2025) develop an automated student answer scoring system using GloVe-LSTM with hybrid similarity metrics, demonstrating strong agreement with human scores on short answer items in Indonesian. The study is methodologically careful: it reports inter-rater agreement baselines, evaluates against multiple similarity metrics, and discusses the gap between technical performance and instructional utility. Yulita, Hariz, et al. (2023) apply deep learning to online exam cheating detection during the COVID-19 pandemic, classifying candidate activity from webcam video alone and reporting an F1-score of about 0.85. While the study addresses a non-language assessment context, its implications for high-stakes remote language testing are direct.

Several contextual studies bear on the assessment pillar indirectly. Indonesian sentiment analysis research (William et al., 2024; Octafia et al., 2026) demonstrates text classification approaches that could support formative analysis of learner writing. Sarcasm detection work (Rosid et al., 2024; Suhartono et al., 2024) addresses the pragmatic dimension of language use, suggesting pathways for assessing learners' interpretive competence beyond literal comprehension though the gap between benchmark performance on social media data and the assessment of learner pragmatic competence is wide, and two findings from this literature are direct warnings for classroom use: fine-tuned language models still outperform large language models on Indonesian sarcasm, with LLMs failing at zero-shot classification entirely (Suhartono et al., 2024), while code-mixed Indonesian–English input requires purpose-built modelling (Rosid et al., 2024) and code-mixing is the ordinary register of Indonesian student talk. Textual entailment recognition (Putra, I. M. S., et al., 2025) provides infrastructure that could support inference-based assessment of reading comprehension, moving assessment beyond literal recall. IndoBERT-based authorship attribution on news text (Saputra & Riccosan, 2024) raises the possibility of stylometric analysis of learner writing development, though the same technology surfaces concerns about authorship verification in AI-assisted writing, and no study has tested whether attribution methods validated on professional journalism transfer to learner corpora.

Each of these implications is speculative. None has been empirically tested in classroom assessment. Together they sketch the contours of a potentially rich multimodal, multi-construct assessment ecosystem for Indonesian language education a sketch, not a finding.

Real-Time Feedback and Multimodal Assessment Possibilities

(Heffernan & Heffernan C. L, 2014) demonstrated through the ASSISTments ecosystem that automated assessment can blend with learning in ways that support rather than replace teacher judgment, providing a precedent for learner-centered AI assessment that the Indonesian literature has yet to engage.The pedagogical promise of deep-learning-based assessment lies in three directions. First, automated scoring of constructed responses could provide real-time formative feedback to learners and reduce teacher workload, particularly in the large classes characteristic of Indonesian secondary and tertiary education. Second, multimodal assessment pipelines could expand what is assessable beyond traditional written products to include oral presentations, multimodal compositions, and interactional performance. Third, fine-grained classification of learner text by sentiment, pragmatic function, or stylistic profile could yield diagnostic information that supports targeted instruction.

Shute (2008), in her review of formative feedback, established that effective feedback is timely, specific, and actionable criteria that automated scoring systems must meet to claim pedagogical rather than merely technical value. Timeliness matters as much as accuracy. SLA research has long held that feedback is most effective when it occurs within the window of cognitive engagement with a task, and that delayed feedback degrades both its uptake and its motivational impact. In Indonesian classrooms, where teachers routinely manage large classes across multiple sections, the latency of feedback on constructed response tasks often extends to days or weeks, well beyond the optimal window. Automated scoring that returns feedback within minutes could therefore represent a genuinely pedagogical advance, provided that the feedback is calibrated not merely to a score but to actionable next steps for revision. The literature in the corpus reports scores but rarely reports feedback content, leaving this pedagogically critical dimension unexamined.

Fairness, Integrity, and the Generative AI Challenge

Five critical concerns dominate.

First, construct underrepresentation is pervasive: models trained on textual features may score surface fluency rather than communicative competence, particularly when training data is drawn from proficient rather than learner writing. Second, fairness across language varieties is underexamined: scoring models trained on standard Indonesian may penalize learners whose writing reflects regional language influence, code-switching, or non-standard but communicatively effective registers. Third, the academic integrity applications raise a pedagogical dilemma: surveillance-oriented cheating detection frames learners as objects of suspicion rather than agents of learning, potentially raising the affective filter that Section 4.1 identifies as antithetical to acquisition. Fourth, the rise of generative AI tools that produce fluent text with minimal learner effort threatens the validity of any assessment that equates text quality with learning.

The generative challenge deserves elaboration. The same Transformer-based technologies that power automated scoring also power generative tools that produce fluent Indonesian text in response to a prompt. The implication is that any assessment based on out-of-class writing is now vulnerable to AI assistance that current detection methods cannot reliably identify, and even in-class writing may be mediated by AI tools accessible through personal devices. Responses to this challenge surveillance, oral defense, in-class writing, process-based assessment each carry pedagogical costs and benefits that the field has only begun to articulate. The Indonesian assessment literature in the corpus largely predates the widespread classroom availability of generative AI; the next wave of research will need to address how formative and summative language assessment can be reconceived where fluent text is no longer a reliable proxy for learning.

Fifth is differential impact across learner populations. Scoring models trained predominantly on data from urban learners with strong exposure to standard Indonesian may perform less accurately on writing from learners whose first language is a regional or non-Austronesian language, or whose writing reflects distinctive code-mixing patterns. The effect would be to penalize precisely the learners whose educational contexts have historically been underserved, reproducing existing inequalities under the description of objective assessment. This remains a hypothesis rather than a documented effect, and that is itself the problem: no study in the corpus reports subgroup performance analyses, so the question cannot currently be answered either way. A pedagogically responsible assessment AI would require not only subgroup validation but also mechanisms for teacher override, learner appeal, and ongoing audit of differential performance none of which are visible in the current literature.

Discriminative versus Generative AI: Pedagogical Implications

The reviewed studies span two orientations of deep learning that carry distinct pedagogical implications.

Discriminative models including IndoBERT, CNN, BiLSTM, and the Transformer-based classifiers that dominate the corpus learn to map inputs to predefined labels: a piece of learner writing is classified as anxious or not, a short answer is scored against a rubric, an utterance is identified as Javanese or Sundanese. Their pedagogical contribution lies in automating detection, scoring, and classification tasks that teachers currently perform manually, potentially freeing teacher attention for higher-order instructional work. Their risk, as documented across Sections 4.1 to 4.3, is that of reducing complex learning phenomena to modelable surface features and of displacing teacher interpretive judgment with algorithmic output.

Generative models large language models capable of producing novel text, speech, and multimodal content operate on a different logic. Their pedagogical affordances include the generation of abundant comprehensible input at adjustable difficulty (Krashen, 1985), the creation of conversational practice partners that can negotiate meaning (Long, 1996), and the production of instant feedback on learner output (Swain, 2005). These affordances align directly with the SLA principles introduced in Section 2.2 in ways that discriminative classifiers cannot. The risks are correspondingly different: generative models can produce fluent text without learning, undermining the construct validity of writing assessment (Section 4.3.4); they can hallucinate culturally inappropriate or factually wrong content; and they can erode learner authorship when used to produce rather than to support learner work.

The corpus reflects a marked asymmetry. As reported in Table 1, 119 of the 134 studies (89%) develop or evaluate discriminative models, while only 15 engage with generative AI, and most of those 15 do so incidentally rather than as the primary object of study. Of the 134 studies, 118 (88%) report only engineering metrics without engaging learning outcomes; 24 studies demonstrate technical capability transferable to language learning; and only 7 frame their contribution in SLA or pedagogical terms at all. This asymmetry reflects the historical trajectory of Indonesian NLP, which has matured through classification and detection tasks. The pedagogical implication is that the field has developed strong infrastructure for assessment and monitoring but weak infrastructure for the generative practices input provision, conversational scaffolding, feedback generation that SLA theory identifies as central to acquisition. Put bluntly, Indonesian AI-in-language-education research has built the tools for watching learners well before it built the tools for teaching them. The learner-centered agenda proposed in Section 5.3 therefore distinguishes between discriminative and generative tools, calibrating evaluation criteria and governance arrangements to the distinct pedagogical risks and affordances of each.

Critical Discussion: A Pedagogical Lens

Cross-Cutting Synthesis

Figure 2. Cross-Pillar Synthesis: Convergence and Tension across the Three Pillars

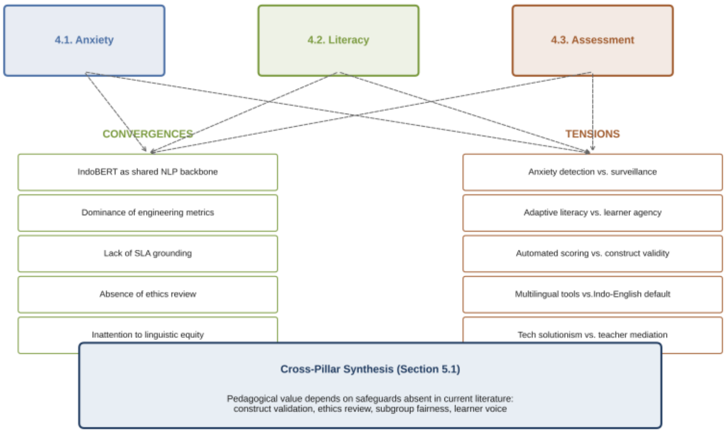


Figure 6. Cross-pillar synthesis: convergences and tensions across the three pillars

Source: Author's analysis based on reviewed literature.

Figure 2 maps convergences and tensions across the three pillars. Technologically, Indonesian deep learning has advanced rapidly with IndoBERT supporting diverse tasks. Pedagogically, the literature shares three limitations: optimization to engineering metrics (only 16 of 134 studies measure learning outcomes), rare SLA/educational theory grounding (only 7 studies), and absent ethics/equity considerations (no ethics review, only six consent procedures, no subgroup analyses).

These limitations reflect global patterns but take specific forms in Indonesia: framing regional languages as "low-resource," assuming Bahasa Indonesia and English as defaults, ignoring the rural urban digital divide, and excluding teacher voice in tool design all reproducing structural inequalities.

Cross-pillar tensions emerge: affective monitoring promises to lower anxiety, while academic integrity surveillance threatens to raise it. Deploying both creates a doubly surveilled environment where any deviation is read as pathology or fraud. A coherent pedagogical approach must address these tensions proactively.

This section proposes a five-criterion SLA framework: input quality, interaction authenticity, output opportunity, affective filter, and learner agency. Unlike edtech frameworks measuring integration degree (TPACK, SAMR, PICRAT), this specifies acquisition-relevant conditions. Since 89% of the corpus uses discriminative models, applying it to generative LLMs requires explicit adaptation.

Five commitments for Indonesia-responsive AI: multilingual by design (NusaX/NusaCrowd), local-context aware (large classes, limited connectivity), ethical and participatory (privacy, consent, community involvement), governed under a concrete learner data regime (school-level assessments, age-appropriate consent, on-device processing, access rights, retention protocols), and teacher AI collaboration.

Realizing these requires sustained teacher education investment across the Kemendikti Saintek–Kemendikdasmen institutional split. Teacher preparation must include critical AI literacy evaluating tools pedagogically and ethically, interpreting AI outputs, and co-designing tools. Without this, AI tools will remain ad hoc and pedagogically unsound.

CONCLUSION

This review of 134 Scopus-indexed publications on deep learning applications in Indonesian language education revealed a significant disconnect between technological advancement and pedagogical implementation. The literature demonstrated rapid technical progress through models such as IndoBERT and related architectures; however, it remained overwhelmingly focused on engineering metrics, with 89% of studies reporting only technical performance measures without evaluating learning outcomes. Furthermore, only 7 of the 134 studies incorporated explicit pedagogical or second-language acquisition (SLA) perspectives, and ethical considerations remained limited, as none of the reviewed studies reported formal ethics review procedures and only six documented consent procedures.

The review made three main contributions. Empirically, it provided a pedagogical mapping of deep learning applications in Indonesian language education. Conceptually, it proposed a five-criterion SLA-informed evaluation framework consisting of input quality, interaction authenticity, output opportunity, affective filter, and learner agency. Methodologically, it highlighted the importance of distinguishing between deep learning as a neural network-based computational approach and pembelajaran mendalam as a pedagogical concept in Indonesian educational discourse. Four priorities for future research were identified: ethnographic studies of classroom implementation, experimental research measuring actual learning outcomes, community-driven development of regional language corpora, and the establishment of a national ethical AI framework for education through multi-stakeholder collaboration. This review concluded that effective AI integration in Indonesian language education depends not only on technological sophistication but also on the development of strong pedagogical and ethical frameworks supported by collaboration among researchers, educators, and policymakers.

REFERENCES

Aji Winata G. I. Koto F. Cahyawijaya S. Romadhony A. Mahendra R. Kurniawan K. Moeljadi D. Prasojo R. E. Baldwin T. Lau J. H. & Ruder S, A. F. (2022). *One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia*. https://doi.org/10.18653/v1/2022.acl-long.500

Anwar & Harmanto B, K. (2025). *Leveraging large language models to detect academic anxiety in Indonesian English for Specific Purposes students through reflective writing*. https://doi.org/10.26803/ijlter.24.12.17

Apandi Mutiara A. B. & Dharmayanti D, A. (2025). *Generating image captions in Indonesian using a deep learning approach based on Vision Transformer and IndoBERT architectures*. https://doi.org/10.47738/jads.v6i2.672

Atika Dwijayanti S. & Suprapto B. Y, R. (2026). *Improving speech-to-text for the Indonesian language using a modified Transformer*. https://doi.org/10.15587/1729-4061.2026.350949

Black & Wiliam D, P. (1998). *Assessment and classroom learning*. https://doi.org/10.1080/0969595980050102

Cahyawijaya Lovenia H. Aji A. F. Winata G. I. Wilie B. Koto F. Mahendra R. Wibisono C. Romadhony A. Vincentio K. Santoso J. Moeljadi D. Wirawan C. Hudi F. Wicaksono M.

S. Parmonangan I. H. Alfina I. Putra I. F. Rahmadani S. … Purwarianti A, S. (2023). *NusaCrowd: Open source initiative for Indonesian NLP resources*. https://doi.org/10.18653/v1/2023.findings-acl.868

Carroll Garba I. Figueroa-Rodríguez O. L. Holbrook J. Lovett R. Materechera S. Parsons M. Raseroka K. Rodriguez-Lonebear D. Rowe R. Sara R. Walker J. D. Anderson J. & Hudson M, S. R. (2020). *The CARE Principles for Indigenous Data Governance*. https://doi.org/10.5334/dsj-2020-043

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research*.

Chen Zou D. Xie H. & Su Y, X. (2021). *Twenty years of personalized language learning research: A bibliometric analysis*.

Enrique Alfina I. & Yulianti E, G. (2024). *Javanese part-of-speech tagging using cross-lingual transfer learning*. https://doi.org/10.11591/ijai.v13.i3.pp3498-3509

Godwin-Jones, R. (2019). *Riding the digital wilds: Learner autonomy and informal language learning*.

Gonçalves de Bem, L., Oliveira Filho, D., Imbuzeiro, H. M. A., Pereira Oliveira, J. V., Motoike, S. Y., Maciel da Costa, V., & Filho, D. L. (2022). Solar photovoltaic tree multi aspects analysis − a review. *Renewable Energy and Environmental Sustainability*, *7*. https://doi.org/10.1051/rees/2022014

Greenhalgh Thorne S. & Malterud K, T. (2018). *Time to challenge the spurious hierarchy of systematic over narrative reviews? European Journal of Clinical Investigation, 48(6), e12931*. https://doi.org/10.1111/eci.12931

Hastuti Hidayat E. Y. Salam A. & Sudibyo U, K. (2026). *KerisRDNet: Mask-aware augmentation and residual dilated networks for cultural heritage blade classification*. https://doi.org/10.1016/j.mlwa.2026.100852

Heffernan & Heffernan C. L, N. T. (2014). *The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching*. https://doi.org/10.1007/s40593-014-0024-x

Holmes, J., & Wilson, N. (2017). *An introduction to sociolinguistics* (5th ed.). Routledge.

Horwitz Horwitz M. B. & Cope J, E. K. (1986). *Foreign language classroom anxiety*. https://doi.org/10.1111/j.1540-4781.1986.tb05256.x

Kalantzis & Cope B, M. (2012). *Literacies*.

Kemendikbudristek. (2022). Standar kompentensi Lulusan Pendidikan Paud, Pendidikan Dasar dan Pendidikan Menengah. *Journal of Chemical Information and Modeling*.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*.

Lhaksmana Falif M. S. Nurhayati I. K. Rezaldi M. Y. Prakasa E. & Roedavan R, K. M. (2026). *Detecting language anxiety in Indonesian students: Deep learning and traditional classification methods for English learning anxiety*. https://doi.org/10.29207/resti.v10i1.7109

Mas Diyasa Idhom M. Saputra A. S. A. Dewi D. A. & Fahrudin T. M, I. G. S. (2025). *Automated student answer scoring using GloVe-LSTM and hybrid similarity metrics*. https://doi.org/10.28945/5629

Muhathir N. Barus R. K. I. Ula M. & Sahputra I, K. (2023). *Preserving cultural heritage through AI: Developing LeNet architecture for wayang image classification*. https://doi.org/10.14569/IJACSA.2023.0140919

Purnomo, A. M., Siti, R., Fauziah, P., Roestamy, M., Martin, A. Y., Adri, H. T., Noor, M. N., & Tirta, R. (2025). Gender, Region, and Field of Expertise in Vocational Education: A Map of Inequality among Vocational High School Students. *ASEAN Journal of Educational Research and Technology*, *4*(3), 334.