REVIEW ARTICLE

Generative AI and the Nuance of *Panitikang Filipino*: A Narrative Review of Technology-Mediated Teaching in Philippine Language Education

Ivy Joy Ganancial[1,*]
[1] University of the Immaculate Conception, Philippines

*Corresponding author: iganancial@uic.edu.ph

Keywords: academic integrity, cultural preservation, digital literacy, e-learning, translation

DOI: https://doi.org/10.66074/X1Y4Z7A5B8

Received: 10 June 2026 Accepted: 27 Aug 2026
Revised: 12 Aug 2026 Published: 30 Aug 2026

ABSTRACT

Background: Generative artificial intelligence has entered language classrooms faster than the evidence base has matured, with most published work centered on English and second-language learning. Filipino language and literature instruction presents a distinct problem because linguistic form, code-switching, historical reference, and culturally embedded meaning can expose weaknesses in large language models. *Objective*: This narrative review examined evidence relevant to the use of generative AI, digital translation tools, and e-learning platforms in Filipino instruction, with emphasis on academic writing, reading comprehension, digital literacy, academic integrity, and culturo-linguistic preservation. *Methods*: Literature published from 2019 to 2025 was identified through structured searches of multidisciplinary education databases, publisher platforms, UNESCO resources, and the ACL Anthology. Evidence was organized by pedagogical function, learner outcome, teacher competence, integrity risk, and Filipino-specific linguistic or cultural performance. *Results*: GenAI research reported frequent use for content generation, feedback, explanation, and writing support, but direct evidence from Filipino private-school classrooms was absent. Filipino benchmarks documented substantial limitations in cultural knowledge, reading comprehension, translation, morphology, and generation. Philippine digital-learning studies showed both instructional potential and persistent access, design, and teacher-readiness constraints. *Conclusion*: GenAI can support Filipino instruction when teachers treat model output as material for critique, revision, and comparison rather than as authoritative language evidence. Filipino-specific validation, transparent AI-use rules, and culturally accountable teacher mediation remain essential.

I. INTRODUCTION

Generative artificial intelligence (GenAI) has altered the practical conditions under which students read, compose, translate, summarize, and revise academic text. Large language models can produce fluent explanations, adapt texts to requested levels, generate examples, offer feedback, and sustain dialogue at a speed that conventional educational technologies rarely match. Reviews of early educational use describe substantial interest in personalization, content generation, formative feedback, tutoring, and teacher support, alongside concern about factual error, bias, overdependence, privacy, and academic misconduct (Farrokhnia et al., 2024; Lo, 2023; Tlili et al., 2023; Wang et al., 2024). The language-education literature reflects the same pattern. Kohnke et al. (2023) identified immediate possibilities for language practice and instructional support, while Yang and Li (2024), in a review of 44 studies on ChatGPT for second-language learning, found that content generation, feedback, and teaching support were dominant uses and that writing skills and learner perceptions were the most common outcomes. This rapidly expanding evidence base has normalized the question of whether GenAI belongs in language education, but it has not answered the more discipline-specific question of what responsible use requires when the language of instruction is itself underrepresented in model development.

The imbalance is especially consequential for Filipino. Filipino and Tagalog are not simply lower-volume versions of English input. They contain morphosyntactic structures, voice and focus systems, flexible word order, reduplication, affixation, pragmatic particles, code-switching practices, and lexical choices whose meanings depend heavily on social and cultural context. Recent Filipino natural language processing research has made these limitations measurable. Montalan et al. (2025) developed Batayan to test understanding, reasoning, and generation in Filipino and Taglish and reported marked performance gaps associated with underrepresentation in pretraining data and the language's morphological and constructional complexity. Miranda et al. (2025) evaluated 27 models through FilBench and found continuing weakness in reading comprehension and translation, with the strongest evaluated model reaching 72.23% overall. Earlier, Montalan et al. (2024) showed an even sharper cultural gap in Kalahi: the best tested model answered 46.0% of culturally nuanced prompts correctly, compared with 89.10% for native Filipino respondents. These results make linguistic fluency an insufficient proxy for linguistic and cultural adequacy.

The issue becomes more demanding in *Panitikang Filipino* because literary study relies on more than sentence-level accuracy. Interpretation requires attention to historical location, social relations, rhetorical indirection, humor, figurative language, regional reference, oral and written traditions, and the cultural significance of particular lexical choices. A model may produce grammatically plausible Filipino yet flatten a character's social position,

normalize a culturally marked expression into generic prose, mistranslate a metaphor, or supply an interpretation that sounds authoritative without adequate textual basis. Cross-cultural research on large language models offers a broader basis for this concern. Tao et al. (2024) documented systematic cultural alignment patterns that often tracked English-speaking and Protestant European cultural profiles, while Blodgett et al. (2020) argued that language technologies embed social assumptions through data, task design, and evaluation choices. Bender et al. (2021) similarly warned that large-scale language modeling can reproduce dominant patterns without grounded understanding. In a Filipino literature classroom, such weaknesses can affect not only correctness but also what counts as a legitimate reading of culture.

At the same time, technology-mediated Filipino instruction did not begin with GenAI. Philippine schools have long used learning management systems, digital modules, online discussion spaces, multimedia, machine translation, and electronic reading materials, although adoption accelerated sharply during the pandemic. Collado et al. (2021), from responses of 86 Filipino teachers in public and private institutions, documented both professional growth and substantial difficulty during the shift to online instruction. Barrot et al. (2021) found that Filipino college students' online-learning difficulties included problems tied to the home learning environment and the overall quality of the learning experience. In language education specifically, Tarrayo et al. (2023) reported that 38 Filipino university English teachers viewed flexible learning as both useful and problematic, with technology, pedagogy, and institutional conditions shaping implementation. Evidence from mother-tongue literacy also shows that well-designed digital materials can support reading development: Semilla et al. (2023) reported positive effects of e-learning materials on early reading skills in a Philippine mother-tongue context. GenAI therefore enters an existing digital ecosystem rather than an instructional vacuum.

What remains underdeveloped is the connection among these bodies of evidence. Broad GenAI scholarship often assumes English-dominant educational contexts, while Filipino NLP benchmarks evaluate model capability without directly examining classroom pedagogy. Philippine digital-learning studies document access, adaptation, and teacher experience but largely predate the widespread classroom use of generative models. Yusuf et al. (2024), who mapped 407 GenAI publications, identified limited K-12 research, limited experimental evidence, and insufficient attention to cultural dimensions. Yang and Li (2024) likewise showed that L2 evidence was concentrated in English-as-a-foreign-language higher education and in studies with relatively small samples. Within this distribution, direct evidence on private-school teachers who use GenAI to teach Filipino language and literature, and on the resulting effects on Filipino academic writing and reading comprehension, remains notably scarce.

A private-school focus is analytically important because institutional conditions may shape both opportunity and risk. Private schools vary widely in connectivity, device access, learning-management infrastructure, teacher professional development, assessment policy, and curricular autonomy. These conditions can determine whether AI becomes a carefully supervised literacy tool, an unexamined shortcut, or an additional source of inequity. Teacher competence is central to this distinction. Ng et al. (2021) conceptualized AI literacy through knowledge and understanding, use, evaluation, and ethical awareness, while Miao and Cukurova (2024) placed human-centered judgment, AI ethics, foundations and applications, AI pedagogy, and professional learning within a teacher competency framework. For Filipino teachers, these competencies require an added language-specific layer: the ability to detect culturally implausible output, check translation choices, distinguish formal Filipino from contextually inappropriate standardization, and preserve students' authorship during AI-assisted writing.

This review therefore examined how GenAI, digital translation tools, and e-learning platforms can be situated within Filipino language and literature instruction, with particular attention to private-school practice. It addressed four connected questions: what pedagogical functions are documented for AI-assisted language teaching; what evidence exists regarding academic writing and reading-comprehension outcomes; what forms of digital and AI literacy are required for responsible use; and how academic integrity and culturo-linguistic preservation constrain instructional design. Rather than assuming that results from English instruction transfer directly to Filipino, the review placed educational studies beside Filipino-specific computational evidence and Philippine digital-learning research. This approach allows the instructional promise of GenAI to be assessed against the actual linguistic and cultural limits of the models students and teachers are likely to use.

II. METHODS

Review Design

This study used a structured narrative review design to integrate evidence from educational technology, second-language pedagogy, Philippine digital learning, AI literacy, academic integrity, and Filipino natural language processing. A narrative approach was selected because the evidence base is methodologically heterogeneous: it includes systematic and rapid reviews, surveys, qualitative studies, mixed-methods research, quasi-experimental work, policy and competency frameworks, and computational language-model benchmarks with non-comparable outcome metrics. A statistical meta-analysis would therefore combine constructs, populations, and performance measures that do not share a common effect metric. The review nevertheless followed explicit procedures to reduce selection and interpretive bias. Methodological decisions were informed by the principles for rigorous evidence synthesis described by Siddaway et al. (2019), while reporting and critical appraisal were guided by the Scale for the Assessment of Narrative Review Articles, or SANRA, which emphasizes justification of the review's importance, statement of aims, description of literature search, appropriate referencing, scientific reasoning, and presentation of relevant endpoint data (Baethge et al., 2019).

Search Strategy and Eligibility

The search covered literature published from January 2019 through December 2025, with the final verification of bibliographic details completed on August 29, 2026. The 2019 start point captured the recent digital-learning period while retaining methodological sources required for review design. Searches were conducted across Google Scholar and ERIC for education literature, the ACL Anthology for Filipino and multilingual NLP research, and publisher or indexing platforms associated with Elsevier, Springer Nature, Wiley, Sage, Taylor & Francis, UNESCO, and relevant open-access journals. Search strings combined technology terms with language, educational, and outcome terms. Core combinations included (generative AI OR ChatGPT OR large language model OR machine translation) AND (Filipino OR Tagalog OR Philippine language OR mother tongue) AND (teaching OR language education OR literature OR writing OR reading OR comprehension); (AI literacy OR digital literacy) AND (teacher OR student) AND (Philippines OR language education); and (academic integrity OR assessment) AND (generative AI OR ChatGPT). Additional searches paired e-learning, flexible learning, translation, cultural bias, morphology, code-switching, Filipino NLP, reading comprehension, private school, and teacher competence. Reference lists of highly relevant reviews and Filipino benchmark papers were used for backward checking when they identified sources within the date and scope criteria.

Evidence Selection

Eligible sources met at least one of five criteria: they examined GenAI or large language models in education or language learning; they examined digital or e-learning conditions in Philippine education with direct relevance to language instruction; they evaluated Filipino, Tagalog, Taglish, or Philippine-language performance of language technologies; they addressed AI literacy, academic integrity, assessment, or ethical use in educational settings; or they provided an authoritative international framework directly relevant to teacher AI competence. Peer-reviewed journal articles, peer-reviewed conference papers from established computational-linguistics venues, and UNESCO guidance or competency frameworks were eligible. Sources were excluded when they were commercial product pages, news reports, undated commentary, non-scholarly blog posts, studies unrelated to education or language, duplicate reports of the same evidence, or papers whose findings could not be verified from a publisher, indexing record, or stable scholarly repository. Opinion pieces were not used to establish outcome claims. Older literature was retained only when it supplied a necessary review-method foundation. Evidence specifically about private-school Filipino instruction was flagged during screening; the absence of direct studies in this category was recorded as a result rather than filled through extrapolation.

Data Extraction and Synthesis

For each included source, the review extracted publication year, country or linguistic context, educational level, participant or dataset characteristics, technology examined, pedagogical function, reported learner or teacher outcome, integrity or ethical concern, and any linguistic or cultural finding. Computational benchmark studies were additionally coded for language coverage, task family, comparison model, and the reported performance statistic most relevant to classroom use. Philippine digital-learning studies were coded for institutional context, modality, teacher or student experience, and barriers to implementation. The evidence was then organized into five analytic domains: AI-assisted pedagogy; digital literacy, translation, and e-learning; academic writing and reading comprehension; academic integrity and assessment; and culturo-linguistic preservation. A sixth cross-cutting code captured private-school relevance. The synthesis used constant comparison across domains rather than vote counting. Direct Filipino evidence received priority when a claim concerned Filipino language capability; Philippine classroom studies received priority for local implementation conditions; language-learning reviews informed pedagogical functions and outcomes; and general AI-education reviews or ethical frameworks were used only for claims that extended beyond Filipino-specific evidence.

Quality and Interpretive Safeguards

Several safeguards were used to maintain evidential discipline. First, claims about model capability were tied to named Filipino benchmarks rather than to anecdotal demonstrations of fluent output. Second, educational outcome claims were separated from tool-affordance claims: a model's ability to generate feedback, for example, was not treated as evidence that the feedback improves Filipino writing. Third, causal language was reserved for studies with designs that supported causal inference; descriptive, survey, qualitative, and review findings were reported as associations, experiences, documented patterns, or reported effects. Fourth, evidence from English or general L2 education was labeled as adjacent evidence and was not presented as direct proof of effectiveness in Filipino. Fifth, cultural claims required either Filipino-specific evaluation or cross-cultural language-model evidence. Finally, the synthesis was checked against SANRA domains after drafting to ensure that the review rationale, search process, endpoint data, and scientific argument were explicit. These procedures produced a review that is reproducible at the level appropriate to a narrative synthesis while remaining transparent about the scarcity of direct classroom evidence for GenAI-assisted Filipino instruction.

III. RESULTS

Scope and Distribution of the Evidence

The reviewed evidence was distributed across four principal literatures: GenAI in education and second-language learning, Philippine digital and flexible learning, Filipino NLP and language-model evaluation, and AI literacy or academic-integrity scholarship. The largest body of evidence concerned general or English-dominant GenAI use. Yusuf et al. (2024) mapped 407 publications and reported concentration in pedagogical enhancement, writing assistance, adoption, ethical risk, and perceptions, with comparatively little K-12 research, experimental work, or culturally focused analysis. Yang and Li (2024) reviewed 44 studies of ChatGPT in L2 learning and reported that most samples involved English-as-a-foreign-language college learners; writing and learner perceptions were more frequently examined than reading or culturally situated language performance. Lo (2023) likewise found that the early educational evidence base was dominated by higher education and that the measured effects varied by task and context.

Philippine classroom evidence was smaller and mostly addressed the transition to digital or flexible instruction rather than GenAI itself. Collado et al. (2021) surveyed 86 teachers from public and private Philippine institutions and reported professional-growth opportunities alongside substantial instructional and technological challenges. Tarrayo et al. (2023) studied 38 Filipino university English teachers through survey and focus-group data and documented both advantages and problems in flexible language teaching. Barrot et al. (2021) reported multiple online-learning challenges among Filipino college students, with the home learning environment among the most prominent. Semilla et al. (2023) provided direct Philippine evidence on technology-supported mother-tongue literacy, reporting improved early reading performance after the use of developed e-learning materials. No identified study directly estimated the effect of GenAI-assisted Filipino teaching by private-school teachers on students' Filipino academic writing and reading comprehension.

Table 1
Representative Evidence Relevant to Filipino Language Education and Generative AI

Source	Context/ design	Technology/ domain	Key reported result
Yang & Li (2024)	Systematic review, 44 L2 studies	ChatGPT in L2 learning	Content generation, feedback, and teaching support were common; writing and learner perceptions dominated; most contexts were EFL higher education.
Yusuf et al. (2024)	Systematic mapping review, 407 publications	GenAI in education and research	Research clustered around pedagogical enhancement, writing, adoption, ethics, and perceptions; K-12, experiments, and cultural dimensions were limited.
Collado et al. (2021)	Survey, 86 Philippine teachers	Transition to online teaching	Teachers from public and private institutions reported professional-growth opportunities and substantial challenges during the shift to cyberspace.
Tarrayo et al. (2023)	Survey and focus groups, 38 Filipino English teachers	Flexible language learning	Teachers reported advantages, disadvantages, and implementation problems tied to flexible instruction.
Semilla et al. (2023)	Philippine mother-tongue e-learning study	Early reading e-learning materials	Contextualized digital materials produced improved early reading performance in the target mother-tongue setting.
Montalan et al. (2024)	Filipino cultural benchmark, 150 prompts	Cultural LLM evaluation	Best model: 46.0% correct; native Filipino respondents: 89.10%.
Montalan et al. (2025)	Filipino and Taglish NLP benchmark	Understanding, reasoning, generation	Performance gaps were reported across tasks, with underrepresentation, morphology, and linguistic construction identified as key difficulties.
Miranda et al. (2025)	27-model Filipino benchmark	Culture, NLP, reading, generation	Best overall score: 72.23%; highest reported Southeast Asian model: 61.07%; reading comprehension and translation remained difficult.
Aquino et al. (2025)	Large-scale Tagalog treebank	Syntactic annotation and parsing	The project documented language-specific annotation and parser challenges associated with Tagalog grammar.
Ng et al. (2021)	Exploratory review	AI literacy	AI literacy was organized around knowledge and understanding, use, evaluation, and ethical issues.
Moorhouse et al. (2023)	Review of top-50 university guidance	GenAI and assessment	Fewer than half had public GenAI assessment guidance; common areas were integrity, assessment design, and communication.
Tao et al. (2024)	Cross-cultural LLM evaluation	Cultural bias and alignment	Model outputs showed systematic cultural alignment patterns; culture-specific prompting improved alignment in many country settings.

Note. The table summarizes representative sources and does not imply equivalence among educational studies and computational benchmarks.

AI-Assisted Pedagogical Functions

Across education and L2 studies, GenAI was used most frequently for content generation, explanation, feedback, practice, lesson support, and revision assistance. Kohnke et al. (2023) described language-learning applications that included conversational practice, idea generation, vocabulary and grammar support, lesson material development, and feedback. Yang and Li (2024) found content generation, feedback, and teaching support to be the most prevalent roles across the 44 L2 studies in their review. Farrokhnia et al. (2024) reported personalization, rapid response, information access, and reduced routine workload among ChatGPT's educational strengths, while also identifying unreliable output, bias, dependency, and integrity threats among its weaknesses and risks. Wang et al. (2024) grouped GenAI affordances in education around accessibility, personalization, automation, and interactivity and identified recurring challenges that included academic integrity, inaccurate or biased output, overreliance, digital inequality, and privacy or security concerns.

The evidence specific to Filipino did not evaluate classroom pedagogy, but it established constraints on the content that GenAI can supply. Montalan et al. (2025) reported performance gaps across Filipino understanding,

reasoning, and generation tasks in Batayan, including Tagalog and Taglish. Miranda et al. (2025) reported weaknesses in Filipino reading comprehension and translation across FilBench. Montalan et al. (2024) found large differences between model and native-speaker performance on culturally nuanced Filipino prompts. The pedagogical functions documented in general language education therefore coexist with measured language-specific limitations in the Filipino context.

Academic Writing and Language Production

Writing was one of the most frequently investigated outcomes in the L2 GenAI literature. Yang and Li (2024) reported that writing skills, together with learner perceptions, formed the dominant objectives across the selected L2 studies. Kohnke et al. (2023) documented uses of ChatGPT for brainstorming, drafting support, language practice, correction, and feedback. Educational reviews also reported frequent use of GenAI for text production, summarization, editing, and feedback (Lo, 2023; Yusuf et al., 2024). These studies primarily involved English or EFL contexts rather than Filipino academic writing.

Filipino computational studies documented uneven performance on the language operations that underlie AI-assisted writing. Batayan evaluated understanding, reasoning, and generation using Filipino and Taglish data and reported persistent gaps across tested models (Montalan et al., 2025). FilBench evaluated 27 models across Filipino-centered tasks, and the highest overall score was 72.23%; the highest-performing Southeast Asian model in the reported comparison scored 61.07% (Miranda et al., 2025). Miranda (2023) also described the need for dedicated Tagalog NLP resources through the development of calamanCy, while Aquino et al. (2025) reported challenges in large-scale Tagalog syntactic annotation and parsing that arose from language-specific grammatical structure. No reviewed classroom study reported a controlled comparison of Filipino academic-writing quality between AI-assisted and non-AI instruction.

Reading Comprehension, Translation, and Digital Text

The reviewed L2 GenAI literature contained less direct outcome evidence for reading comprehension than for writing. Yang and Li (2024) identified writing and perceptions as the most common study objectives, while broader reviews catalogued summarization, question generation, explanation, and adaptive content as common educational uses (Lo, 2023; Wang et al., 2024). In the Philippines, Semilla et al. (2023) reported gains in early mother-tongue reading skills after the use of contextualized e-learning materials, providing evidence that digital delivery can support native-language literacy when content is designed for the target language and learner level. Barrot et al. (2021), however, documented environmental and experiential barriers that affected online learning conditions among Filipino students.

FilBench supplied direct evidence on language-model reading and translation performance. Miranda et al. (2025) reported that several evaluated models had difficulty with reading-comprehension and translation tasks in Filipino, Tagalog, and Cebuano. Batayan likewise reported performance gaps tied to Filipino morphology, construction, and limited representation in pretraining data (Montalan et al., 2025). These benchmark studies used computational task scores rather than student-learning outcomes. No reviewed study tested whether AI-generated Filipino summaries, translations, or comprehension questions improved students' comprehension of *Panitikang Filipino* relative to teacher-authored materials.

Digital Literacy, Teacher Competence, and Platform Conditions

AI literacy frameworks consistently included technical knowledge, critical evaluation, use, and ethics. Ng et al. (2021) synthesized AI literacy into four broad aspects: knowledge and understanding, use, evaluation, and ethical issues. Miao and Cukurova (2024) described 15 teacher competencies across a human-centered mindset, ethics of AI, AI foundations and applications, AI pedagogy, and AI for professional learning, with progression from acquisition to deeper and more creative levels of competence. Miao and Holmes (2023) also emphasized human agency, data privacy, age-appropriate use, and validation of GenAI in education.

Philippine studies documented variable conditions for technology-mediated teaching. Collado et al. (2021) found that Filipino teachers viewed the transition to cyberspace as both a professional opportunity and a demanding instructional adjustment. Tarrayo et al. (2023) reported problems as well as advantages among English-language teachers who shifted to flexible learning. Barrot et al. (2021) identified home-environment and learning-quality challenges among Filipino students. Across these sources, digital access alone did not define readiness; teacher adaptation, task design, institutional support, and learner conditions were recurrent components of technology use. The reviewed literature did not provide a validated measure designed specifically for Filipino teachers' ability to evaluate GenAI output in Filipino or *Panitikang Filipino*.

Table 2
Evidence Pattern by Review Domain

Domain	Documented affordances	Documented constraints	Filipino-specific evidence status
AI-assisted pedagogy	Idea generation; explanations; feedback; practice; lesson and question generation; revision support	Hallucination; bias; overreliance; weak transfer evidence to Filipino	Direct classroom evidence absent; Filipino benchmarks establish capability constraints
Digital literacy and e-learning	Flexible access; digital texts; contextualized materials; professional learning	Access inequity; teacher preparedness; task-design demands; privacy	Philippine digital-learning evidence available; Filipino-specific AI-literacy measure not identified
Academic writing	Brainstorming; drafting support; feedback; language revision	Authorship ambiguity; inaccurate correction; homogenized style; unsupported claims	Writing evidence mainly EFL; no controlled Filipino private-school outcome study identified
Reading comprehension	Summaries; explanations; adaptive questions; vocabulary support	Interpretive flattening; inaccurate inference; weak source grounding	FilBench reports reading difficulty; Philippine mother-tongue e-learning evidence is adjacent rather than GenAI-specific
Translation	Rapid comparison across languages; lexical alternatives; scaffolds for discussion	Literalization; loss of register, idiom, pragmatics, and cultural reference	FilBench reports persistent Filipino translation difficulty
Academic integrity	Process feedback; transparent assisted revision when permitted	Unauthorized generation; attribution problems; validity of unsupervised assessment	General guidance is substantial; Filipino-language detection evidence was not identified
Culturo-linguistic preservation	Comparison of interpretations; explicit critique of model choices; local corpus work	Cultural bias; standardization; erasure of regional or contextual nuance	Kalahi, Batayan, FilBench, and Tagalog NLP work document substantial representation gaps

Note. Evidence status describes the literature identified in this review through December 2025.

Academic Integrity and Assessment

Academic integrity appeared as a recurrent concern across GenAI education research. Cotton et al. (2024) described risks related to unauthorized text generation, attribution, and assessment validity and emphasized the difficulty of preserving conventional take-home assessment assumptions after ChatGPT. Perkins (2023) distinguished AI use from misconduct by locating integrity in disclosure, policy, and the nature of the permitted assistance. Moorhouse et al. (2023), in a review of assessment guidance among the world's 50 top-ranked universities, found that fewer than half had publicly available GenAI-related guidance at the time of review; the available guidance centered on academic integrity, assessment design, and communication with students. Farrokhnia et al. (2024) and Wang et al. (2024) also identified integrity, overreliance, and reliability as recurring educational risks.

No reviewed source established a Filipino-language AI-detection method with sufficient validity to serve as a primary integrity mechanism in school assessment. The Filipino benchmark literature instead documented that model outputs can be fluent while remaining limited in culture, comprehension, translation, or language-specific reasoning (Miranda et al., 2025; Montalan et al., 2024, 2025). This evidence separates surface fluency from verified linguistic accuracy and from evidence of student authorship.

Culturo-Linguistic Representation and Preservation

The strongest Filipino-specific evidence concerned cultural and linguistic representation. Kalahi used 150 handcrafted prompts created by native Filipino speakers to evaluate culturally appropriate responses. The best tested

model answered 46.0% correctly, compared with 89.10% for native Filipino respondents (Montalan et al., 2024). FilBench included cultural knowledge, classical NLP, reading comprehension, and generation and reported that the best evaluated model achieved an overall score of 72.23% (Miranda et al., 2025). Batayan used native-speaker-driven adaptation and validation across Filipino and Taglish tasks and documented substantial performance gaps related to underrepresentation, morphology, and linguistic construction (Montalan et al., 2025). Aquino et al. (2025), through the UD-NewsCrawl Treebank, likewise documented annotation and parsing challenges that arose from distinctive Tagalog syntax.

Cross-cultural evidence showed comparable problems at a broader scale. Tao et al. (2024) found systematic cultural bias and alignment patterns in large language models and reported that culture-specific prompting improved alignment in many country contexts but did not remove the underlying variation. Blodgett et al. (2020) documented how NLP systems and evaluations can encode social and institutional assumptions, while Bender et al. (2021) described risks that arise when large language models reproduce statistical regularities from massive corpora without grounded knowledge of social meaning. Across the reviewed evidence, no benchmark score established parity between current LLMs and Filipino native-speaker judgment across linguistic, cultural, literary, and pragmatic dimensions.

IV. DISCUSSION

The central implication of this review is that the educational question for *Panitikang Filipino* is not whether GenAI can produce Filipino text, but whether a particular use preserves disciplinary judgment. Fluency is now cheap and immediate, whereas culturally credible interpretation remains difficult to automate. The Filipino benchmarks make that distinction unusually visible. Kalahi showed a large gap between native-speaker and model performance on culturally grounded prompts (Montalan et al., 2024), while Batayan linked persistent model difficulty to Filipino morphology, construction, and limited representation in pretraining data (Montalan et al., 2025). FilBench added direct evidence that reading comprehension and translation remain challenging even for strong contemporary models (Miranda et al., 2025). These findings change the pedagogical role that should be assigned to AI in Filipino classes. A generated paragraph is best treated as a contestable text that students must inspect, compare, correct, and contextualize, not as a reference answer whose surface confidence substitutes for language expertise.

This distinction is especially important in literary instruction. A literature classroom asks students to make warranted interpretations from textual evidence, to recognize ambiguity, and to relate language choices to history, identity, power, and community. GenAI can assist with preparatory tasks such as vocabulary exploration, question generation, alternative paraphrases, or the comparison of possible readings. Yet the cultural evidence reviewed here indicates that these outputs require teacher-controlled verification. Tao et al. (2024) showed that model responses can align systematically with dominant cultural profiles, while Blodgett et al. (2020) demonstrated why bias in language technology cannot be reduced to a few offensive outputs: it also appears in what data are represented, what tasks are treated as normal, and whose language practices define correctness. Bender et al. (2021) further cautioned that statistical language generation does not provide grounded social understanding. In *Panitikang Filipino*, a responsible task can exploit this weakness pedagogically by asking students to identify what an AI interpretation omits, whose perspective it privileges, and which phrases lose cultural force after paraphrase or translation.

Academic writing offers a different but related opportunity. The L2 literature already documents GenAI use for brainstorming, feedback, drafting support, and revision (Kohnke et al., 2023; Yang & Li, 2024). These functions can reduce the cost of iterative practice, especially when a teacher has limited time for line-by-line feedback. However, Filipino academic writing requires more than transferring an English writing workflow into a Filipino prompt. Model feedback itself may contain questionable lexical choices, overly formalized constructions, literal translations, or stylistic convergence toward generic prose. The safest instructional design is therefore scaffolded comparison: students first produce a traceable draft, request a defined form of AI assistance, compare the response with dictionaries, course texts, teacher feedback, and peer judgment, then justify which suggestions they accepted or rejected. This keeps the student responsible for the final linguistic decision. It also aligns with Ng et al.'s (2021) view of AI literacy as not only use but also evaluation and ethics, and with Miao and Cukurova's (2024) emphasis on human-centered teacher competence.

For reading comprehension, the review supports a similarly bounded role. GenAI can generate pre-reading questions, explain unfamiliar vocabulary, offer alternative summaries, and produce prompts at different difficulty levels, but these features are affordances rather than proof of learning. Wang et al. (2024) described personalization and interactivity among the recurring educational affordances of GenAI, yet FilBench shows that model performance on Filipino reading tasks is itself imperfect (Miranda et al., 2025). This means that an AI-generated comprehension question can be pedagogically useful only after someone has checked that the question is answerable from the text, that its key is defensible, and that it does not simplify a culturally or rhetorically complex passage into one supposedly correct interpretation. Semilla et al. (2023) provides an instructive contrast: technology-supported mother-tongue reading can improve performance when materials are deliberately developed and contextualized for the target learners. The important variable is not digitality by itself, but the quality of language-specific instructional design.

Digital translation deserves particular caution because it can create an illusion of equivalence. Translation tools and LLMs are useful for displaying alternatives quickly and can help students compare how lexical choice changes tone or meaning. In a Filipino literature course, that capacity is strongest when translation becomes an object of analysis. Students can compare a human translation, a model translation, and the original, then annotate shifts in register, metaphor, politeness, social relation, or cultural reference. The direct benchmark evidence supports this critical stance: Miranda et al. (2025) found translation to remain a difficult Filipino task across evaluated models, and Montalan et al. (2025) documented broader limits tied to language-specific structure. Such tasks convert a technical limitation into metalinguistic work without requiring the teacher to treat the machine translation as linguistically authoritative.

The review also clarifies what digital literacy should mean in a native-language context. Conventional digital competence often focuses on access, navigation, communication, information evaluation, and content creation. GenAI adds another layer because the learner must evaluate a system that can respond persuasively even when its evidence or language choice is weak. Ng et al. (2021) identified evaluation and ethical awareness as core AI-literacy dimensions, and the UNESCO teacher framework adds human agency, ethics, foundational knowledge, pedagogy, and professional learning (Miao & Cukurova, 2024). For Filipino teachers, these competencies should include the ability to test a prompt across variants, recognize code-switching and register errors, check culturally specific claims against authoritative sources, inspect whether a translation preserves pragmatic force, and explain why a fluent response may still be linguistically inferior to a student or expert alternative. This is a form of disciplinary AI literacy, because evaluation criteria come from Filipino language and literature rather than from generic technology proficiency.

Academic integrity cannot be separated from that literacy. Cotton et al. (2024) show why unsupervised text production destabilizes familiar assumptions about take-home writing, while Perkins (2023) places emphasis on policy clarity and disclosure instead of treating all AI use as identical misconduct. Moorhouse et al. (2023) found that early university guidance clustered around integrity, assessment redesign, and communication with students. These principles can be adapted to Filipino classes through process-visible assessment. A teacher can require planning notes, annotated sources, draft histories, oral explanation of key language choices, a short AI-use declaration when assistance is permitted, and a comparison between original and AI-revised passages. Tasks based on local texts, classroom discussion, or student-specific interpretation also make authorship more visible. Such designs assess the reasoning behind the Filipino prose, not merely the polish of the final surface form, and they avoid dependence on AI-detection tools that have not been validated for Filipino classroom use.

Culturo-linguistic preservation should not be framed as resistance to technology. The stronger position is to make local language knowledge a criterion that technology must meet. Filipino NLP research already demonstrates the value of native-speaker-built resources. Kalahi was created from culturally nuanced prompts by Filipino speakers (Montalan et al., 2024); Batayan used native-speaker-driven adaptation and validation to avoid translationese and to represent Filipino and Taglish more faithfully (Montalan

et al., 2025); and Aquino et al. (2025) exposed structural challenges through large-scale Tagalog syntactic annotation rather than forcing the language into assumptions derived from better-resourced languages. Classroom practice can follow the same logic. Teachers can curate local texts, retain code-switching when it carries social meaning, invite comparison across regional expressions, and reject AI revisions that erase voice for the sake of standardized fluency. Preservation, in this sense, is active linguistic judgment rather than technological abstinence.

The private-school context deserves direct empirical study rather than indirect generalization from the literature reviewed here. Collado et al. (2021) included teachers from public and private Philippine institutions, but the study concerned the wider transition to online teaching and did not isolate Filipino language pedagogy or GenAI. Tarrayo et al. (2023) offers detailed evidence on flexible English-language teaching, yet its state-university context and English focus limit direct transfer. A rigorous next study could recruit multiple private secondary schools with different levels of digital infrastructure and compare carefully specified AI-assisted and teacher-led Filipino lessons across a full grading period. Outcomes should include independently scored Filipino academic writing, passage-based reading comprehension, retention, source use, student AI literacy, and documented integrity events. Teacher-level measures should include language-specific AI evaluation skill, instructional use logs, and policy knowledge. Such a design would directly test whether carefully specified GenAI use affects Filipino writing and reading outcomes rather than presuming that general GenAI benefits already apply to Filipino.

A useful practice model emerging from the synthesis is a pedagogy-verification-preservation cycle. Pedagogy begins with a learning objective that would remain educationally valuable even if no AI tool were available. Verification requires the teacher and students to check factual claims, textual evidence, lexical choice, translation, register, and cultural fit before accepting model output. Preservation requires a final decision about whether the AI-assisted product retains the linguistic variety, cultural meaning, authorship, and interpretive responsibility expected in the course. The cycle is compatible with the human-centered principles in UNESCO guidance (Miao & Holmes, 2023) and with the teacher competency framework of Miao and Cukurova (2024), but it adds a language-specific criterion demanded by Filipino benchmark evidence. It also avoids the common error of defining innovation as the mere presence of a generative tool.

V. LIMITATIONS

Several limits qualify the conclusions. A narrative review can integrate heterogeneous evidence more effectively than a narrowly defined meta-analysis, but it does not provide a pooled effect size and may miss relevant studies despite structured searching. The direct empirical literature on GenAI in Filipino school instruction is still sparse, so parts of the pedagogical argument rely on adjacent evidence from English or L2 learning, Philippine flexible learning, and Filipino NLP benchmarks. Benchmark scores are also model- and version-specific; they can change as systems, prompts, and training data change. The review window ended in 2025 to preserve a defined evidence period, even though model development continues rapidly. These limits do not support either wholesale adoption or rejection. They narrow the defensible claim: GenAI has plausible instructional value in Filipino education, but effectiveness, equity, and cultural adequacy still require direct classroom validation.

VI. CONCLUSION

Generative AI has created genuine new possibilities for Filipino language and literature instruction, but the available evidence does not justify treating those possibilities as established effects in private-school Filipino classrooms. Research in English and L2 education documents frequent use of GenAI for feedback, explanation, content generation, revision, and writing support. Philippine studies show that digital platforms can widen access to instructional resources and can support native-language literacy when materials are deliberately contextualized. At the same time, Filipino-specific benchmarks reveal substantial weaknesses in cultural knowledge, reading comprehension, translation, morphology, reasoning, and generation. These strands of evidence converge on a bounded role for AI: it can provide material for learning, but it should not define linguistic correctness or literary meaning on its own.

For *Panitikang Filipino*, the most defensible form of integration places teacher expertise, student authorship, source evidence, and cultural judgment above model fluency. AI-assisted writing should preserve visible student decision-making; AI-assisted reading should use verified prompts and text-grounded interpretation; translation should become an exercise in comparison rather than unquestioned substitution; and assessment should make the process of composition and AI use transparent. The next empirical priority is a multi-site study of private-school Filipino classrooms that measures academic writing, reading comprehension, AI literacy, integrity, and cultural-linguistic quality under clearly defined instructional conditions. Until such evidence exists, responsible practice depends on pedagogy first, verification throughout, and preservation of Filipino linguistic and cultural nuance as a non-negotiable outcome.

Author Contributions

The author conceptualized the review, developed the search strategy, synthesized the evidence, interpreted the findings, prepared the language support framework, and approved the final manuscript.

Funding

No external funding was received for this work.

Ethics Approval and Informed Consent

Ethics approval and informed consent were not required because this study reviewed and synthesized previously published literature and publicly available curriculum materials and did not involve direct participation of human subjects.

Competing Interests

The author declares no competing interests.

Data Availability

All evidence discussed in this review is available in the cited publications and official sources. The structured evidence matrix used for synthesis can be made available by the author upon reasonable request.

Declaration of AI Use

A generative artificial intelligence tool was used to assist with language refinement, structural organization, and consistency checking during manuscript preparation. Scholarly claims and bibliographic details were checked against source records, and responsibility for the final content rests with the author.

Acknowledgments

None.

References

Aquino, A. A., Miranda, L. J. V., & Or, E. M. T. (2025). The UD-NewsCrawl Treebank: Reflections and challenges from a large-scale Tagalog syntactic annotation project. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7219-7239). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.357

Baethge, C., Goldbeck-Wood, S., & Mertens, S. (2019). SANRA-a scale for the quality assessment of narrative review articles. *Research Integrity and Peer Review, 4*, Article 5. https://doi.org/10.1186/s41073-019-0064-8

Barrot, J. S., Llenares, I. I., & del Rosario, L. S. (2021). Students' online learning challenges during the pandemic and how they cope with them: The case of the Philippines. *Education and Information Technologies, 26*(6), 7321-7338. https://doi.org/10.1007/s10639-021-10589-x

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of bias in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454-5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Collado, Z. C., Concha, C. B. A., & Orozco, N. M.-I. G. (2021). Teaching in transition: How do Filipino teachers face the migration to cyberspace amid the pandemic? *Computers in the Schools, 38*(4), 281-299. https://doi.org/10.1080/07380569.2021.1988317

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

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International, 61*(3), 460-474. https://doi.org/10.1080/14703297.2023.2195846

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*(3), 504-526. https://doi.org/10.1007/s40593-021-00239-1

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

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), Article 410. https://doi.org/10.3390/educsci13040410

Miao, F., & Cukurova, M. (2024). AI competency framework for teachers. UNESCO.

Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO.

Miranda, L. J. V. (2023). calamanCy: A Tagalog natural language processing toolkit. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)* (pp. 1-7). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.nlposs-1.1

Miranda, L. J. V., Aco, E., Manuel, C. G., Cruz, J. C. B., & Imperial, J. M. (2025). FilBench: Can LLMs understand and generate Filipino? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 2496-2529). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.127

Montalan, J. R., Layacan, J. P., Africa, D. D., Flores, R. I. S., Lopez, M. T., II, Magsajo, T. D., Cayabyab, A., & Tjhi, W. C. (2025). Batayan: A Filipino NLP benchmark for evaluating large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 31239-31273). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1509

Montalan, J. R., Ngui, J. G., Leong, W. Q., Susanto, Y., Rengarajan, H., Aji, A. F., & Tjhi, W. C. (2024). Kalahi: A handcrafted, grassroots cultural LLM evaluation suite for Filipino. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation* (pp. 497-523). Tokyo University of Foreign Studies.

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open, 5*, Article 100151. https://doi.org/10.1016/j.caeo.2023.100151

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology, 58*(1), 504-509. https://doi.org/10.1002/pra2.487

Perkins, M. (2023). Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice, 20*(2), Article 7. https://doi.org/10.53761/1.20.02.07

Semilla, J.-R. B., Parmisana, V. R., Fajardo, L. L., Abucayon, R. L., Dinoro, A. P., & Tabudlong, J. M. (2023). Innovation in early reading instruction: The development of e-learning materials in mother tongue. *International Journal of Learning, Teaching and Educational Research, 22*(7), 411-433. https://doi.org/10.26803/ijlter.22.7.22

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology, 70*, 747-770. https://doi.org/10.1146/annurev-psych-010418-102803

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus, 3*(9), Article pgae346. https://doi.org/10.1093/pnasnexus/pgae346

Tarrayo, V. N., Paz, R. M. O., & Gepila, E. C., Jr. (2023). The shift to flexible learning amidst the pandemic: The case of English language teachers in a Philippine state university. *Innovation in Language Learning and Teaching, 17*(1), 130-143. https://doi.org/10.1080/17501229.2021.1944163

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*, Article 15. https://doi.org/10.1186/s40561-023-00237-x

Wang, N., Wang, X., & Su, Y.-S. (2024). Critical analysis of technological affordances, challenges and future directions of generative AI in education: A systematic review. *Asia Pacific Journal of Education, 44*(1), 139-155. https://doi.org/10.1080/02188791.2024.2305156

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

Yusuf, A., Pervin, N., Román-González, M., & Noor, N. M. (2024). Generative AI in education and research: A systematic mapping review. *Review of Education, 12*(2), Article e3489. https://doi.org/10.1002/rev3.3489