Indonesian Journal of Education Research (IJoER)
Vol. 7, No. 4, August 2026, pp. 463~479
ISSN: 2722-1326, DOI: 10.37251/ijoer.v7i4.2954

A Documentary Evaluation of Offline AI Integration in Teacher-in-a-Box OER Programmes for Low-Connectivity Schools

Kate Huth[1], Jason Zagami[2]
[1,2] School of Education and Professional Studies, Griffith University, Brisbane South (Nathan) Campus, Australia

Article Info

Article history:

Received Mar 30, 2026
Revised Jun 25, 2026
Accepted Aug 7, 2026
Online First Aug 26, 2026

Keywords:

Artificial Intelligence
Assessment Literacy
Character Education
Offline AI
Open Educational Resources
Teacher in a Box

ABSTRACT

Purpose of the study: To conduct a documentary, ex ante programme evaluation of how a bounded offline AI augmentation could be specified, governed and later tested within an offline OER teacher development programme delivered through Teacher in a Box.

Methodology: Documentary programme evaluation using a thesis corpus. Methods: Theory of Change reconstruction, CIPP evaluation framework, directed qualitative content analysis with a scope-aligned codebook, and construct-to-instrument mapping for assessment tools. Offline AI specification is GPT4All class local LLMs plus retrieval-augmented generation (RAG) over the local Teacher in a Box OER and training repository, with governance wrappers.

Main Findings: Offline OER effectiveness depends on integrated, in-situ teacher support rather than access alone. Offline AI is most defensible as a governed, retrieval-grounded, teacher-facing layer that improves resource discovery, supports formative assessment routines, and scaffolds assessment literacy. The study outputs measurable CIPP indicators, an evidence-to-implication warrant structure, a baseline-versus-augmentation risk and safeguard matrix, and offline-feasible instruments for formative assessment, summative task quality, character education evidence, and deeper learning portfolios.

Novelty/Originality of this study: This study provides an audit-ready, evaluation methodology linking offline OER programmes to offline AI. It advances knowledge by separating evidence, implication, and design claims; specifying a minimum viable governance package; and producing offline-feasible assessment instruments and indicators for formative, summative, character, and deeper learning assessment under low-connectivity conditions, supporting policy and implementation decisions before field trials.

Corresponding Author:
Jason Zagami
School of Education and Professional Studies, Griffith University, Brisbane South (Nathan) Campus, 170 Kessels Rd, Nathan QLD 4111, Australia
Email: j.zagami@griffith.edu.au

1. INTRODUCTION

Open Educational Resources (OER) are frequently advanced as a response to inequitable access to learning materials, yet in low-connectivity environments, the conditions of use often determine whether OER can be adopted, adapted, and sustained [1]-[4]. The study underpinning this paper documents an "offline internet" approach in which a local server, Teacher in a Box (TIB), a volunteer-run not-for-profit project/initiative [5], stores resources and makes them available within a school via local Wi-Fi, addressing the practical limitation that most OER ecosystems presume reliable internet access [1], [3]. This further indicates that access alone does not reliably translate into classroom uptake: teachers require targeted, integrated support

to locate, adapt, and integrate OER, particularly where educators have uneven experience with educational technologies and limited time for self-directed exploration [1], [4], [6].

These constraints are especially salient in Indonesia, where archipelagic scale, uneven digital infrastructure, and extensive linguistic diversity shape the feasibility of technology-mediated reform at scale. UNESCO has described Indonesia as a large and diverse archipelago of more than 17,000 islands and around 700 spoken languages and has noted that uneven digital infrastructure remains a significant educational challenge [7], [8]. In 2023, the implementation of Kurikulum Merdeka was explicitly structured around school readiness and curricular flexibility, with schools permitted to implement the curriculum according to their local preparedness [9], while UNESCO reported that Indonesia's public digital learning platforms have positioned digital platforms and technology-supported learning as mechanisms for enhancing educational equity and quality [7], [10]. However, reform also introduces recurring system costs and distribution risks, including the practical burden of accessing nominally free digital resources where effective use still depends on paid or unreliable connectivity [11]. Offline repositories can stabilise access and enable localisation, but their educational value depends on whether teachers can readily find, adapt, and use resources within ordinary classroom conditions rather than merely having the files stored locally [5], [11], [12].

Assessment and evaluation are pivotal to this operationalisation problem. When curriculum aims move towards competencies, practical learning, and student development, assessment regimes must evolve to maintain constructive alignment between policy intent and classroom practice [13]. Formative assessment practices require teachers to elicit evidence, interpret it, and respond instructionally in ways that improve learning, rather than merely generating scores [14]-[16]. In parallel, educators' assessment literacy shapes the quality of both formative and summative assessment: teachers must translate curriculum aims into evidence-rich tasks, apply criteria consistently, and use evidence to guide next instructional steps [17]-[19]. This positions teacher learning as the mechanism by which an offline platform and integrated training programme can improve classroom practice and student outcomes [1], [20]. This paper extends that logic to assessment, treating assessment capability-building and instrument development as core programme functions rather than downstream implementation details.

Character education introduces additional assessment challenges. Many systems seek to promote values, dispositions, and pro-social behaviours alongside academic competencies, but assessment approaches for character education remain contested and vulnerable to superficial measurement, social desirability bias, and cultural mismatch [19], [21]-[23].

Where connectivity is limited, assessment solutions must also remain feasible without dependence on cloud platforms, analytics dashboards, or high-bandwidth submissions. These pressures create a design space in which programme-level interventions must support both curriculum implementation and assessment capability-building with tools that are robust under constrained conditions. Recent advances in generative AI suggest a potential contribution to this design space: local-first language models that can operate offline after initial setup (for example, GPT4All class of tools) and can be paired with retrieval methods that ground outputs in a locally stored corpus [24]-[27]. For offline OER programmes, the central question is not whether AI can replace teacher judgement, but whether a constrained, governed offline AI layer can strengthen programme functions: improving local resource discoverability, supporting formative assessment routines, scaffolding educator assessment literacy, and assisting in the development of learning instruments and media [28], [29]. This is conceptually congruent with the study's emphasis on repackaging dispersed training materials into integrated, bite-sized supports embedded within an offline server to improve usability at the point of need [1], [6]. In effect, offline AI can be treated as an interface to those supports, making them searchable, context-responsive, and actionable without requiring connectivity.

At the same time, introducing AI into curriculum and assessment work raises risks that must be evaluated rather than assumed away. Risks include over-standardisation of judgement, amplification of cultural mismatch, unreliable guidance, and "assessment gaming" in which students optimise for outputs rather than learning [25], [26], [28], [30], [31]. These risks intensify when assessment targets deeper learning outcomes and character development, where evidence is multi-source, context-dependent, and easily distorted by reductive indicators [19], [21]-[23]. A postcolonial framing is therefore not a background narrative but an evaluative constraint: it centres equity and power relations in educational development and highlights the risks of externally imposed assumptions and dominant narratives shaping what counts as legitimate knowledge, practice, and success within intervention settings [1]. In the context of offline AI, this implies that governance and localisation are not optional features. They are the conditions under which offline AI either supports educational sovereignty or reintroduces dependency through imported assumptions embedded in models, prompts, and criteria [26], [28], [30].

A further reason to examine offline AI separately from cloud AI is that the technical architecture changes the educational and policy problem. Cloud-based systems presume dependable connectivity, ongoing subscription costs, external data processing, and platform governance that is usually controlled far from the school or ministry using the system [26], [28], [30]. By contrast, a local-first configuration running on an offline

device within the school can be bounded more tightly around a known corpus, known users, and known update cycles. This does not remove bias or validity risk, but it changes what can be governed locally: storage, prompt libraries, model updates, language resources, and approval workflows can all be treated as programme components rather than invisible backend dependencies [24]-[26], [28], [32]. For low-connectivity schools, this matters because educational sovereignty is not only about having access to content, but about having practical control over how support tools behave, what knowledge they privilege, and what data leaves the site. In assessment terms, the distinction is equally important. A cloud service may be attractive for scale, but an offline system can be designed to keep draft work, rubric exemplars, moderation notes, and teacher queries inside the local environment, reducing unnecessary data exposure while keeping support available at the point of need [17], [26], [28], [30].

This focus also helps clarify why the paper centres on teacher development rather than AI capability alone. Assessment is not improved merely because a system can generate feedback, a quiz, or a rubric. Improvement depends on whether teachers can judge the fit between curriculum intent, evidence, task design, and learner context, and whether the programme infrastructure helps them make those judgements more reliably over time [13], [16]-[19]. In that sense, offline AI should be understood as one component within a broader socio-technical support system that includes curated OER, training artefacts, localisation, school routines, and moderation practices. The most consequential question is therefore not whether the model is impressive in isolation, but whether it strengthens the practical chain linking access, teacher learning, classroom adaptation, and trustworthy assessment under real low-connectivity conditions [1], [6], [13], [17], [20], [32].

Existing OER research has established the importance of access, licensing, reuse, adaptation and teacher uptake, yet it has less often specified how offline repositories can become evaluable teacher development systems under low-connectivity conditions. At the same time, much AI in education research assumes cloud-based platforms, persistent connectivity, data-intensive infrastructures and institutional governance capacity that may not be available in remote or resource-constrained schools. The unresolved problem is thus not simply whether AI can be added to OER, but how an offline AI augmentation can be specified, bounded and evaluated before implementation so that programme claims remain proportionate to the available evidence

While offline OER platforms and offline servers have been studied as access solutions in constrained contexts [1], [2], evaluation-ready specifications for AI-augmented offline OER programmes remain underdeveloped. In particular, there is a need for (i) programme evaluation artefacts that clarify how offline AI modifies programme inputs and processes, (ii) indicators that are meaningful for curriculum functions and policy impact, and (iii) offline-feasible assessment instruments and media for formative assessment, summative task quality, educator assessment literacy, character education, and deeper learning outcomes. This paper addresses that need through a documentary, ex ante programme evaluation that makes baseline assumptions explicit and models the implications of an offline AI augmentation under clear governance constraints. The gap is not only the absence of an evaluation of AI for OER. It is the absence of an evaluation-ready framework for judging whether offline AI can be added to an offline OER programme without overstating impact, weakening teacher judgment, importing external assessment norms, or creating unsustainable technical dependencies. Previous work on OER and offline repositories has shown that access infrastructure is necessary, but has less often specified how embedded support, assessment literacy and governance should be evaluated as part of the programme logic. Conversely, recent AI in education research has generated important debates about feedback, assessment, personalisation and governance, but much of this work assumes internet-connected systems and platform infrastructures that do not fit low-connectivity schools [33]-[36]. This study addresses that gap by treating offline AI as a proposed augmentation to be evaluated conceptually and programmatically before field deployment.

Research aim is loosely specified AI integration risks producing claims that exceed the available evidence. The aim of this study is to conduct a documentary, ex ante programme evaluation of how a bounded offline AI augmentation could plausibly reshape curriculum support, assessment practice and educator assessment literacy within an offline OER programme, and to identify the safeguards and evidence requirements needed before field implementation. The research questions (RQs) in this study are:

RQ1. What curriculum functions, roles, and objectives are evidenced in an offline OER programme model, and how would an offline AI augmentation plausibly alter these functions and associated policy impacts?

RQ2. What implications does offline AI have for classroom assessment practices (formative and summative) and educator assessment literacy within an offline OER programme?

RQ3. What constructs, evidence sources, and safeguards are required to assess character education credibly when offline AI tools are available to teachers and students in low-connectivity contexts?

RQ4. What new assessment forms and instruments are needed to assess deeper learning outcomes and student development in competency-based curricula supported by offline OER and offline AI?

RQ5. What boundary conditions would need to be met before any component of the offline AI augmentation model could be considered for adaptation in other ASEAN contexts?

This paper contributes an audit-ready evaluation methodology that extends an established offline OER intervention logic into the assessment and evaluation domain. It provides: (i) a reconstructed programme Theory of Change and a CIPP evaluation frame; (ii) an implication matrix comparing baseline offline OER with an offline AI augmentation specification; and (iii) draft, offline-feasible assessment instruments and indicators for formative assessment, summative task quality, educator assessment literacy, character education, and deeper learning outcomes. The study does not claim that offline AI has improved teacher practice, student learning, assessment quality or programme sustainability; it specifies what would need to be true, governed and measured before such claims could responsibly be made.

To avoid conflating evidence with inference, the paper distinguishes three claim types. Evidence claims are grounded in the study's corpus [1]. Implication claims are derived by applying the Theory of Change and CIPP logic to a narrowly defined offline AI augmentation (local model, retrieval grounding, and governance wrapper) [13], [24], [25]. Design claims which take the form of instrument prototypes and validity-oriented requirements derived from assessment scholarship and the constraints identified in the corpus [14]-[22].

This claim discipline matters because ex ante studies can easily slide into technological optimism. In settings where infrastructure is constrained and educational reform is ongoing, the practical consequences of premature design assumptions are substantial: schools may invest scarce time in tools that are hard to maintain, teachers may be asked to trust outputs that have not been locally warranted, and ministries may overestimate the transferability of technical solutions developed elsewhere. By making explicit what is evidenced, what is inferred, and what is proposed, the study aims to support a more careful mode of innovation. Its purpose is not to slow experimentation unnecessarily, but to ensure that experimentation proceeds from a clear understanding of programme logic, local constraints, and the forms of educational evidence that would count as success [1], [13], [25], [37].

It also sharpens the paper's intended use. The manuscript is designed to inform programme design conversations, implementation planning, and the construction of later field studies. Read in that way, its value lies in disciplined preparation for action rather than in claiming that action has already succeeded.

2. RESEARCH METHOD

2.1 Research Design and Rationale

This study contributes to the use of documentary programme evaluation. It conducts an ex ante (forward-looking) evaluation and implication analysis of integrating offline AI tools into an offline OER programme delivered via Teacher in a Box (TIB). The purpose is to generate decision-support artefacts, rather than to estimate causal effects. The evaluation focuses on three priorities: (i) curriculum functions, roles, objectives, and policy implications, (ii) classroom assessment practice and educator assessment literacy (including character education), and (iii) development of offline-feasible assessment instruments and media for low-connectivity settings. Accordingly, the study should be read as an evaluability and design study, not as a field evaluation, impact study, intervention trial or implementation report.

The evaluation approach integrates three complementary components. First, Theory of Change (ToC) reconstruction is used to make the programme logic explicit, including assumed mechanisms and boundary conditions. Second, the CIPP evaluation frame (Context, Input, Process, Product) structures evaluative questions and indicator selection around improvement and decision needs [13], [33]. Third, directed qualitative content analysis is used to code the documentary corpus with a scope-aligned codebook, while allowing inductive refinement of sub-categories as patterns emerge [38].

2.2 Nature of Claims and Inference Control

To strengthen auditability and avoid conflating documentary evidence with forward-looking design reasoning, the paper distinguishes three claim types.

1. Evidence claims are statements grounded directly in the study's corpus and are traceable to specific segments recorded in the claim-evidence table [1].
2. Implication claims are derived by applying ToC and CIPP reasoning to a narrowly defined offline AI augmentation specification. These claims are framed as plausible programme implications under stated design constraints, not as measured effects [25].
3. Design claims are assessment instruments and media specifications justified through assessment scholarship and validity-oriented reasoning, given the constraints identified in the corpus [14], [16]-[19], [32].

This separation is enforced operationally during analysis: each claim entered into the synthesis tables is labelled as evidence, implication, or design, and each implication or design claim is paired with its warrant (ToC mechanism, CIPP domain, and literature-based assessment requirement).

Analytically, the study proceeded in four passes. First, the thesis corpus was read to reconstruct the baseline intervention logic, including the problem setting, intended users, mechanisms of change, and contextual constraints. Second, these baseline statements were organised through a Theory of Change lens and then mapped into CIPP categories so that programme claims could be translated into evaluation questions and observable indicators [13], [25]. Third, a deliberately narrow offline AI specification was introduced as a hypothetical augmentation, limited to a local language model, retrieval over an approved offline corpus, and governance controls on prompting, access, and review [24], [25]. Fourth, the implications of that augmentation were traced across curriculum support, formative and summative assessment, teacher assessment literacy, character education, and transferability. At each stage, the analysis prioritised conservative reasoning: where the corpus did not support a strong inference, the paper retained the baseline claim and framed the AI implication as conditional rather than asserted.

This procedure was designed to make the paper useful for decision-making before field deployment. Many technology proposals in education move quickly from technical possibility to implementation advocacy, leaving questions of evaluability, validity, and local fit to later stages. The present design reverses that sequence. It first asks what the existing programme is already trying to accomplish, what mechanisms appear to matter, and what evidence would be needed to justify any extension into AI-supported activity. This is especially important when the proposed system may shape teachers' assessment work. In such cases, the quality of an intervention depends not only on whether it can generate outputs, but on whether those outputs are warranted, reviewable, and compatible with the values and capacity of the setting in which they will be used [14], [17], [25], [39].

2.3 Documentary corpus and claim limitations

The empirical corpus comprises an unpublished thesis manuscript reporting the TIB study context, programme design, and evidence regarding OER use in low-resource settings [1]. The thesis defines "offline internet" and positions TIB as a Wi-Fi-enabled local server providing access to stored educational content within a confined area [1]. The thesis also establishes programme-relevant conditions used as boundary constraints in this evaluation, including: (i) access to OER does not reliably translate into classroom uptake without targeted support; (ii) OER professional learning is often dispersed and delivered in large, decontextualised units that are difficult to operationalise; and (iii) a postcolonial framing that centres equity and power relations and foregrounds the risks of externally imposed assumptions shaping what counts as legitimate knowledge, practice, and success [1].

Although the corpus is a single thesis manuscript, it is analytically suitable for the present purpose because the evaluation question is programme-focused rather than population-estimating. The study does not attempt to infer national prevalence or effect sizes. Instead, it examines a documented intervention model in enough depth to reconstruct its operating logic, identify the mechanisms it privileges, and ask how a carefully bounded AI augmentation would interact with those mechanisms. For documentary programme evaluation, the critical issue is therefore not the number of data sources alone, but whether the corpus contains sufficient detail about context, design intentions, supports, and constraints to permit disciplined analysis [13], [25]. The thesis provides that detail by specifying the TIB model, its low-connectivity rationale, the practical barriers to OER uptake, and the importance of integrated teacher support [1]. This is a major boundary on the study's claims. A single thesis corpus does not provide triangulated evidence of teacher practice, classroom use, student outcomes, implementation fidelity, model performance, user acceptance or sustainability. The analysis is thus limited to reconstructing programme logic, identifying plausible implications and specifying evaluation requirements for later field testing.

The unpublished status of the thesis also shaped the study's inferential caution. Because the corpus has not been triangulated here with additional field materials, interview follow-up, or classroom artefacts, the analysis avoids making causal claims about realised impact. Instead, it treats the thesis as a bounded empirical record from which programme mechanisms and constraints can be reconstructed. This makes the paper strongest as an evaluability and design study: it shows what should be monitored, what forms of support appear most defensible, and what validity conditions would need to be met before wider implementation claims could be justified [13], [25], [32]. No new data were collected, and the analysis does not involve interaction with teachers, students, or schools beyond what is already documented in the thesis corpus.

2.4 Unit of Analysis

The unit of analysis is a meaningful documentary segment (for example paragraph, table, figure caption, or described artefact) that specifies:

a) Programme functions, roles, objectives, and mechanisms, including contextual constraints (infrastructure, device access, language ecology, teacher capability);

b) Assessment-related implications (explicit or implicit links to formative practice, summative task design, educator capability, student outcomes); and
c) Policy-relevant claims (scalability, sustainability, equity, governance) [1].

2.5 Evaluation Framework Aligned to Scope

A scope-aligned evaluation matrix was specified a priori and applied as a directed coding frame. The top-level domains correspond to the evaluation foci addressed in this paper.

Domain A: Curriculum evaluation (functions, roles, objectives, and policy impact). Programme functions and objectives inferred from the corpus are mapped to the reconstructed ToC and CIPP frames, then used to model how an offline AI layer could alter roles (teacher, student, programme support), decision points, and policy levers such as equity, quality assurance, and sustainability [25].

Domain B: Classroom assessment and educator assessment literacy (including character education). The analysis identifies how the thesis' logic of integrated training support extends to assessment practice supports (formative routines, feedback cycles, summative task quality) and specifies the additional assessment literacy demands that arise when offline AI is introduced into planning and classroom workflows [14], [17].

Domain C: New constructs and forms for assessing deeper learning outcomes, student development, and test management. The analysis derives feasible constructs for deeper learning assessment that remain implementable offline and identifies integrity and management risks introduced by AI-augmented work processes [14], [17].

Domain D: Learning assessment for development of learning instruments and media. The analysis produces offline-feasible instrument prototypes (rubrics, checklists, portfolio structures) and pairs each with a brief validity-oriented rationale for intended use, evidence requirements, and foreseeable threats.

Domain E: Indonesia and ASEAN transfer logic. Indonesia-specific constraints identified in the corpus are treated as boundary conditions for feasibility and policy impact. Transferability is treated as conditional and is analysed through an explicit boundary-condition rubric specified in Section 2.9.

2.6 Offline AI Augmentation Specification (the Evaluand)

The offline AI component is specified narrowly so it is evaluable and governance-relevant rather than a general claim about generative AI. The augmentation specification includes three elements.

1. A locally hosted language model capable of offline operation after initial setup, treated as an implementation family rather than a single model brand. GPT4All is used as an illustrative reference point because it documents an ecosystem aimed at widening access to local model use [24].
2. A retrieval-augmented generation (RAG) layer that grounds responses in locally stored OER and training materials, prioritising local content sovereignty and reducing unsupported outputs by constraining generation to a curated offline corpus [25].
3. A governance wrapper comprising approved prompt templates, role-based access rules, and quality assurance checks for teacher-facing and student-facing use cases. These governance elements are evaluation requirements rather than assumed features.

Within CIPP, the offline AI augmentation is treated as an evaluable change to programme Inputs and Processes, with downstream effects on Products framed as implications to be tested in subsequent empirical evaluation [25].

2.7 Research Procedure and Reproducibility Controls

The analysis produced a baseline ToC model and an AI-augmented ToC model, a CIPP evaluation matrix with indicators, an implication matrix (baseline vs augmented), and draft assessment instruments with validity notes.

Reproducibility is supported through three controls. First, a claim-evidence table is maintained so that evidence claims and implication claims are traceable to a specific corpus segment, along with their inference rule (ToC mechanism and CIPP domain). Second, the codebook is versioned: top-level scope codes are fixed a priori, while sub-codes are refined inductively and logged with dates and rationales in the analytic memo [32]. Third, negative case analysis is applied during synthesis: segments that constrain optimistic assumptions (for example, feasibility, equity, cultural mismatch, and workload) are recorded explicitly and used to qualify implication claims.

Coding proceeded in four stages. First, all documentary segments relevant to programme function, teacher support, OER use, assessment, governance, equity, infrastructure and sustainability were identified. Second, these segments were coded deductively against the scope-aligned domains and CIPP categories. Third, sub-codes were refined inductively where the corpus identified recurring constraints, such as download cost, language fit, workload, fragmented training materials or cultural mismatch. Fourth, coded segments were translated into claim records, with each record classified as evidence, implication or design. Claims that could

not be traced to a documentary segment, evaluation rule or cited literature warrant were removed or rewritten as future research requirements.

2.8 Minimum Viable Governance Package (Evaluation Requirement)

- Prompt library governance: a versioned library of approved prompts aligned to local curriculum and assessment norms, with templates prioritising evidence-first reasoning.
- Role-based access: differentiated access for teachers, students, and administrators, including restrictions for summative assessment contexts.
- Retrieval scope restriction: responses must be grounded in the local corpus for instructional and assessment-related queries; open-ended generation is restricted or disabled for designated assessment uses [24].
- Quality assurance checks: a structured review process for locally authored exemplars, rubrics, and prompt templates, including periodic bias and mismatch checks.
- Update protocol: a documented process for updating corpus content, retriever index, and model binaries, with rollback capability.
- Integrity protocols: explicit rules for what constitutes acceptable AI assistance in formative and summative contexts, paired with assessment designs that require process evidence and oral verification where feasible.
- Accountability documentation: evaluation reasoning and decision trails are documented to support internal metaevaluation and responsible use, consistent with evaluation standards that emphasise transparency, consequences, and documentation [33].

This package should be treated as part of the evaluand, not as an implementation afterthought. Operational checklist provided in Appendix F. Where these governance requirements cannot be met, the offline AI augmentation should be considered not ready for implementation, regardless of technical feasibility.

2.9 Indonesia and ASEAN Transferability Rubric (Boundary-Condition Logic)

Transferability is assessed through a boundary-condition rubric intended to prevent over-generalisation across ASEAN contexts. The rubric evaluates the plausibility of the offline AI augmentation under seven contextual conditions: (i) device and energy reliability; (ii) local language coverage and script support; (iii) teacher professional learning infrastructure; (iv) assessment policy regime (for example high-stakes dominance versus classroom-based emphasis); (v) local governance capacity for curation and moderation; (vi) procurement and data governance constraints relevant to locally hosted systems; and (vii) curriculum reform churn and update burden [1]. These conditions are used to qualify implication claims and to justify staged rollout recommendations in later sections. The rubric does not demonstrate ASEAN transferability. It identifies the conditions that would need to be examined before transfer could be proposed. No claims are made about effectiveness or feasibility in other ASEAN countries without additional country-specific data.

2.10 Trustworthiness and Evaluation Quality

Trustworthiness and evaluation quality were addressed through analytic transparency, claim discipline and, where possible, expert validation of the evaluation artefacts. The audit trail links each claim to corpus evidence, Theory of Change mechanisms, CIPP domains and warrant rules. Negative case analysis was used to identify feasibility, equity, workload, cultural mismatch, infrastructure and sustainability constraints that limited optimistic AI claims. Because the study did not collect new field data, member checking, classroom observation, interview triangulation and intercoder reliability testing were not conducted. These procedures are identified as requirements for the next empirical phase.

3. RESULTS AND DISCUSSION

3.1 Results

The results are reported as evaluation outputs rather than implementation findings. The study did not test an offline AI intervention in schools, measure teacher uptake, or estimate effects on student learning. Instead, it reconstructed the programme logic of the Teacher in a Box (TIB) offline OER model and examined how a narrowly bounded offline AI augmentation would alter programme inputs, processes, risks, safeguards and future evidence requirements.

3.1.1 Baseline Programme Findings

Finding 1: TIB is framed as an “offline internet” access intervention. The thesis describes TIB as a local server that hosts educational content and makes it accessible via local Wi-Fi within a limited radius, addressing

the reliance of most OER ecosystems on connectivity [1]. This establishes access stabilisation as a primary programme function in low-connectivity settings.

Finding 2: Access is necessary but insufficient; uptake depends on integrated support embedded in the offline environment. The thesis indicates that teachers require targeted, integrated support to locate, adapt and integrate OER effectively, particularly when educators are not regular users of educational technology and have limited time for self-directed exploration [1]. It further motivates repackaging dispersed professional learning into smaller, more usable supports embedded within the server to improve usability at the point of need.

Finding 3: Curriculum reform increases distribution burdens and makes "free" resources practically costly. The thesis notes that during curriculum change, updated materials may be available without purchase cost, yet downloading requires paid connectivity, creating recurring access and distribution burdens at scale. This functions as a policy-relevant justification for offline repositories and local adaptation workflows.

Finding 4: Equity and power are evaluation constraints, not background context. The thesis adopts a postcolonial framing that centres equity and power relations and highlights risks of externally imposed assumptions shaping what counts as legitimate knowledge, practice and success within intervention settings [1]. This establishes a non-negotiable constraint for any augmentation: added tools must not reintroduce dependency through imported norms embedded in systems, prompts and criteria.

Synthesis of baseline ToC: The baseline programme theory implied by the thesis is that stabilised offline access reduces reliance on connectivity, integrated and in-situ teacher support increases the likelihood of uptake, and local availability and adaptation reduce the practical burdens created by curriculum reform churn and download costs. Teacher capability building is the principal mechanism linking access to classroom use and student outcomes [1]. These findings define the baseline programme logic; they do not establish that offline AI improves TIB or teacher practice.

3.1.2 CIPP-by-scope evaluation matrix

Table 1 translates the baseline Theory of Change into decision-oriented indicators and specifies how an offline AI augmentation would modify what should be monitored. Baseline indicators reflect what the thesis implies should be evaluated [1]. AI-augmented indicators reflect additional measurement requirements introduced by the augmentation specification [24], [25].

Table 1. CIPP-by-Scope Evaluation Matrix for TIB and Offline AI Augmentation

CIPP domain	Scope-aligned evaluation focus	Baseline indicators (measurable)	AI-augmented indicators (additional measurable requirements)
Context	Curriculum functions, roles, objectives; policy impact	Connectivity reliability (days/week); device-to-teacher ratio; download cost constraints; teacher edtech confidence distribution	Local governance capacity (named roles, time allocation); language coverage needs (languages used in school); AI access equity (who can use, when)
Input	Learning instruments and media	Presence of curated OER corpus; availability of embedded training supports; format diversity (PDF, video, modules)	Local model deployment feasibility (hardware, power); retrieval corpus defined and bounded; versioned prompt library; offline instrument packs present
Process	Classroom assessment practice; assessment literacy	Teacher onboarding completion; frequency of resource discovery and adaptation; training usage frequency; classroom integration instances	Frequency of AI-assisted discovery/planning use; rate of retrieval-grounded responses (system setting); QA workflow adherence; integrity protocol adherence
Product	Outcomes (formative, summative, character, deeper learning) and policy impact	Teacher uptake (active users/month); reduced reliance on paid downloads; reported usability of supports; sustainability signals (local maintenance)	Assessment routine consistency (formative cycles/week); rubric/task quality improvements (moderation checks); integrity incidents; equity incidents (exclusion, bias reports)

The baseline programme logic implies that evaluation must attend to both infrastructure constraints and the capability layer created by integrated supports [1]. Under CIPP, offline AI is best treated as a change to inputs and processes that generates additional governance and integrity indicators, rather than as an assumed improvement to products [25].

3.1.3 Evidence-to-Implication Mapping

Table 2 provides an explicit warrant structure by linking each implication claim and the evaluation rule used. This structure makes clear that the results include what the thesis establishes and what follows only under the specified augmentation and evaluation rules.

Table 2. Example Claim–Evidence–Warrant Mapping (Audit Excerpt)

Claim ID	Claim Type	Scope Domain	Evidence Anchor (Thesis)	Warrant Rule Applied	Resulting Implication Claim
C1	Evidence	Curriculum function	Offline access needed; Wi-Fi local server model	Not applicable	Access stabilisation is core programme function
C2	Evidence	Capability building	Uptake requires integrated, in-situ supports	Not applicable	Capability layer is central mechanism
C3	Implication	Assessment literacy	Integrated supports improve uptake	If supports drive uptake, then AI as support interface may increase usability (ToC)	Offline AI should be teacher-facing support, not assessor
C4	Implication	Formative assessment	Teachers need targeted supports	Add retrieval-grounded routines as process scaffold (CIPP Process)	AI can scaffold formative routines if bounded to local corpus
C5	Implication	Character education	Equity/power risks foregrounded	Governance must localise constructs and norms (CIPP Input/Process)	Character assessment must use locally endorsed constructs + triangulation
C6	Design	Instruments/media	Integrated supports embedded offline	Assessment theory: validity and evidence sufficiency [14], [17]	Produce offline instruments with validity notes and integrity safeguards

3.1.4 Baseline Versus Offline AI Implication Matrix

The offline AI augmentation is defined narrowly: a locally hosted model capable of offline operation, retrieval grounding to the local corpus, and a governance wrapper of approved prompts, role-based access and QA checks [24], [25]. Table 3 summarises decision-relevant implications. These are not claims of impact; they are evaluable predictions about what would change in programme roles, processes, risks and monitoring requirements.

Table 3. Implication Matrix: Baseline TIB Versus TIB Plus Offline AI Augmentation

Scope area	Baseline claim (evidence)	Implication pathway (under augmentation specification)	Main risks	Required safeguards (minimum viable)
Curriculum functions and policy impact	Offline access + embedded supports enable uptake [1]	AI acts as an interface to embedded supports (discovery, localisation, planning)	Imported norms; uneven access within schools	Locally curated prompt library; role-based access; equity access rules
Formative assessment practice	Uptake requires targeted support [1]	AI provides retrieval-grounded formative routines (hinge questions, exit tickets, feedback prompts)	Unreliable guidance; compliance without instructional response	Evidence-first templates; teacher verification step; retrieval scope locked to corpus
Summative assessment and test management	Reform churn increases burden [1]	AI assists with drafting task variants and rubrics aligned to objectives	Task leakage; gaming; sameness	Controlled access; versioning; moderation routines; integrity protocols
Educator assessment literacy	Teachers need capability building [1]	AI prompts scaffold criteria clarity, evidence sufficiency, fairness checks	Dependency and deskilling	"Judgement support" framing; reflective prompts; exemplars; human-in-the-loop
Character	Equity/power	AI supports locally	Cultural mismatch;	Community-endorsed

education assessment	constraints are central [1]	curated scenarios and reflection prompts anchored to local constructs	performativity	constructs; multi-source evidence rules; moderation guidance
Deeper learning and student development	Competency focus implies authentic evidence [1]	AI supports portfolios and self-assessment prompts grounded in local resources	Ghost-writing; authenticity loss	Process evidence requirements; staged drafts; oral checks; peer critique protocols
Indonesia–ASEAN transferability	Infrastructure and language vary [1]	Value highest where connectivity constraints are high and governance feasible	Transfer failure if governance capacity is weak	Minimum governance package; localisation toolkit; staged rollout

Across domains, the lowest-regret positioning for offline AI is as teacher-facing support for discovery, planning, formative routines and assessment literacy scaffolding. This follows directly from the thesis mechanism: integrated support increases practical uptake [1]. Attempting to position AI as an assessment authority increases integrity, fairness and cultural mismatch risks without corresponding documentary evidence of readiness. Additional implementation risks include uneven local model performance, language mismatch, hardware and power requirements, maintenance burden, teacher deskilling and procurement dependency [40], [41].

3.1.5 Draft Assessment Instruments and Media

These instruments are design outputs for future testing; they have not yet been psychometrically validated or field-tested [37], [38]. Because the baseline programme logic privileges embedded supports, instrument prototypes are designed to be stored and used offline, with or without AI. Each instrument is paired with a brief validity-oriented requirement set.

Table 4. Offline-Feasible Instrument Set (Design Outputs)

Instrument	Intended use	Evidence sources (offline feasible)	AI role (bounded)	Validity threats	Mitigations required
Formative Routine Checklist (FRC)	Improve consistency of formative cycles	Hinge questions, exit tickets, teacher notes, quick quizzes	Generate retrieval-grounded prompt variants; suggest interpretation steps	"Feedback theatre"; shallow compliance	Require recorded responsive action; exemplars; retrieval locked to corpus
Summative Task Quality Rubric (STQR)	Improve task alignment, clarity, feasibility	Task brief, rubric, exemplars, student work samples	Draft criteria language; propose parallel practice items	Task sameness; gaming	Versioned templates; moderation guide; controlled access
Assessment Literacy Self-Audit (ALSA)	Build teacher judgement and fairness checks	Annotated rubrics, lesson plans, reflections	Prompt teacher reasoning ("what evidence would change judgement?")	Over-reliance on AI	Reflection-first workflow; teacher justification fields
Character Education Construct Map + Rubric (CECMR)	Support formative character education assessment	Scenarios, reflection logs, peer feedback, teacher observation	Generate locally curated scenarios and prompts anchored to approved constructs	Cultural bias; performativity	Community endorsement; multi-source evidence rule; moderation
Deeper Learning Portfolio Template (DLPT)	Assess development over time	Drafts, process logs, artefacts, oral checks	Structure portfolio; prompt reflective commentary grounded in corpus	Ghost-writing; authenticity loss	Process evidence; staged drafts; oral verification points

Full instrument templates are provided in Appendix A to E; the governance checklist is provided in Appendix F. These instruments operationalise the paper's central design claim: offline AI can improve usability of embedded supports if it is bounded by retrieval grounding and governance, and if assessment designs require evidence and process, not polished outputs alone [14], [16]-[18], [25], [31].

3.1.6 Indonesia and ASEAN Boundary-Condition Rubric

Regional adaptation is treated as conditional and requiring future comparative evaluation. Based on infrastructure variability, paid or unreliable connectivity, language ecologies and school-readiness differences in Indonesia, implication claims are qualified through a boundary-condition rubric comprising seven conditions: device and energy reliability; local language coverage; teacher professional learning infrastructure; assessment policy regime; governance capacity for local curation and moderation; procurement and data governance constraints; and curriculum reform churn [7], [12], [20], [26], [30].

Under this rubric, a staged rollout is recommended: begin with teacher-facing discovery and planning supports, then expand to formative and assessment literacy scaffolding once local instruments and moderation routines are in place, and only then consider any student-facing uses in assessment contexts. The study provides no direct data from other ASEAN countries.

3.2 Discussion

3.2.1 Draft Assessment Instruments and Media (Offline-Feasible, Validity-Oriented)

Because the baseline programme logic privileges embedded supports, instrument prototypes are designed to be stored and used offline, with or without AI. Each instrument is paired with a brief validity-oriented requirement set, drawing on formative assessment foundations and assessment literacy scholarship [13], [14]. Character education is treated as requiring construct clarity and triangulation due to contestation and cultural sensitivity [17], [21].

Full instrument templates are provided in Appendix A–E; the governance checklist is provided in Appendix F. These instruments operationalise the paper's central design claim: offline AI can improve usability of embedded supports if it is bounded by retrieval grounding and governance, and if assessment designs require evidence and process, not polished outputs alone [14], [16]-[18], [25], [31]. Character education instruments must be locally authored and moderated to honour equity and power constraints [1], [21]-[23].

3.2.2 Indonesia and ASEAN Transferability (Boundary-Condition Rubric)

Transferability is treated as conditional rather than assumed. Based on infrastructure variability, paid or unreliable connectivity, language ecologies, and school-readiness differences in Indonesia, implication claims are qualified through a boundary-condition rubric comprising seven conditions: device and energy reliability; local language coverage; teacher professional learning infrastructure; assessment policy regime (high-stakes dominance versus classroom-based emphasis); governance capacity for local curation and moderation; procurement and data governance constraints relevant to locally hosted systems; and curriculum reform churn [7], [12], [20], [26], [30]. Under this rubric, a staged rollout is recommended: begin with teacher-facing discovery and planning supports, then expand to formative and assessment literacy scaffolding once local instruments and moderation routines are in place, and only then consider any student-facing uses in assessment contexts.

3.2.3 Interpretation and Implications

The evaluation suggests that offline AI is most defensible when treated as infrastructure-aware teacher support rather than automation. This distinction is central. The documentary evidence does not support claims that AI improves learning, assessment or teacher practice. It supports a narrower design implication: where offline OER programmes already rely on embedded teacher support, a governed and retrieval-grounded AI layer may reduce friction in resource discovery, planning, formative routines and assessment literacy work.

This framing also protects teacher judgment. AI-generated prompts, rubrics or feedback should not be treated as authoritative evidence. Their value depends on whether they help teachers reason more clearly about curriculum alignment, evidence sufficiency, fairness, cultural fit and next instructional action. In low-connectivity settings, the future of learning is unlikely to be secured by more tools alone. It will depend on whether tools are governed in ways that strengthen local professional capacity rather than replacing or bypassing it.

For policy and programme design, the implication is that offline AI should be evaluated as a governed support layer. The model, retrieval corpus, prompt library, access rules, update processes, moderation routines and integrity protocols are not technical details. They are part of the intervention. Where these governance requirements cannot be met, offline AI should not be treated as implementation-ready.

3.2.4 Risks, Limitations and Future Research

The risks are material. Local models may perform unevenly, particularly across languages and culturally specific educational contexts. Retrieval grounding can reduce unsupported outputs, but it does not guarantee valid assessment reasoning or cultural fit. Hardware, power, maintenance, model updates and corpus governance may introduce sustainability burdens that are easily underestimated. Student-facing use also raises

assessment integrity risks, including ghost-writing, task leakage and polished outputs that obscure actual competence [37], [38], [42].

The study's limitations follow from its documentary design. The corpus is a single unpublished thesis manuscript, and no new interviews, observations, teacher surveys, classroom artefacts, system logs or student outcome data were collected. Member checking, field triangulation and intercoder reliability testing were not conducted. The findings should thus be interpreted as evaluation design outputs and conditional predictions, not as estimates of impact.

Future research should begin with staged field evaluation of low-risk teacher-facing functions. Initial trials should examine whether offline AI-supported retrieval and formative prompts reduce teacher search time, improve alignment between resources and curriculum objectives, or support more consistent use of assessment evidence. Later studies could examine language localisation, hardware reliability, moderation burden, teacher assessment literacy and the sustainability of local governance routines.

4. CONCLUSION

This documentary, ex ante programme evaluation examined how a bounded offline AI augmentation could be specified for an offline OER teacher development programme delivered through Teacher in a Box. The analysis indicates that offline AI is most defensible as a governed, retrieval-grounded, teacher-facing support layer, rather than as an automated assessment authority.

The study contributes a reconstructed Theory of Change, CIPP-aligned indicators, an evidence-to-implication warrant structure, a risk and safeguard matrix, a boundary-condition rubric, and offline-feasible assessment instruments for future testing. These outputs support programme planning and evaluability; they do not demonstrate implementation effectiveness, teacher uptake, student learning gains or ASEAN transferability.

Future work should begin with staged empirical testing of low-risk teacher-facing functions before any student-facing or assessment-adjacent uses are considered. Such trials should examine language coverage, hardware reliability, governance capacity, assessment integrity, teacher professional learning support and sustainability under local conditions.

ACKNOWLEDGEMENTS

None.

REFERENCES

[1] K. Huth, "Maximising resource usefulness in donated materials for student achievement and teacher education," Ph.D. dissertation, Griffith University, 2026, unpublished.
[2] L. Hosman, C. Walsh, M. Pérez Comisso, and J. Sidman, "Building online skills in off-line realities: The SolarSPELL Initiative (Solar Powered Educational Learning Library)," *First Monday*, vol. 25, no. 7, 2020, doi: 10.5210/fm.v25i7.10839.
[3] UNESCO, *Recommendation on Open Educational Resources (OER)*. Paris, France: UNESCO, 2019.
[4] J. Hilton III, "Open educational resources and college textbook choices: A review of research on efficacy and perceptions," *Educational Technology Research and Development*, vol. 64, pp. 573–590, 2016, doi: 10.1007/s11423-016-9434-9.
[5] Teacher in a Box, "What is Teacher in a Box? The technical stuff," Teacher in a Box, accessed Mar. 2, 2026.
[6] D. Wiley and J. L. Hilton III, "Defining OER-enabled pedagogy," *International Review of Research in Open and Distributed Learning*, vol. 19, no. 4, 2018, doi: 10.19173/irrodl.v19i4.3601.
[7] UNESCO, "UNESCO's digital initiatives promoting linguistic diversity in Indonesia," Feb. 1, 2024.
[8] UNESCO, "Launch of the 2023 UNESCO GEM Report: Transforming education with technology in Southeast Asia," Feb. 28, 2025.
[9] Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, "Perpanjangan jangka waktu pendaftaran Implementasi Kurikulum Merdeka secara mandiri tahun ajaran 2023/2024," *Sistem Informasi Kurikulum Nasional*, 2023.
[10] UNESCO, "Unleashing innovation: Embracing digital transformation in education in Indonesia," Sep. 9, 2024.
[11] Teacher in a Box, "Educational resources," accessed Mar. 24, 2026.
[12] Teacher in a Box, "Our story," accessed Mar. 24, 2026.
[13] D. L. Stufflebeam and C. L. Coryn, *Evaluation Theory, Models, and Applications*, 2nd ed. San Francisco, CA, USA: Jossey-Bass, 2014.
[14] P. Black and D. Wiliam, "Assessment and classroom learning," *Assessment in Education: Principles, Policy & Practice*, vol. 5, no. 1, pp. 7–74, 1998, doi: 10.1080/0969595980050102.
[15] D. Boud and E. Molloy, "Rethinking models of feedback for learning: The challenge of design," *Assessment & Evaluation in Higher Education*, vol. 38, no. 6, pp. 698–712, 2013, doi: 10.1080/02602938.2012.691462.
[16] M. Heritage, *Formative Assessment: Making It Happen in the Classroom*. Thousand Oaks, CA, USA: Corwin, 2010.
[17] W. J. Popham, "Assessment literacy for teachers: Faddish or fundamental?," *Theory Into Practice*, vol. 48, no. 1, pp. 4–11, 2009, doi: 10.1080/00405840802577536.

[18] C. DeLuca, D. LaPointe-McEwan, and U. Luhanga, "Teacher assessment literacy: A three-dimensional model," *Teaching and Teacher Education*, vol. 84, pp. 128–138, 2019, doi: 10.1016/j.tate.2019.05.003.
[19] G. P. Wiggins, *Educative Assessment: Designing Assessments to Inform and Improve Student Performance*. San Francisco, CA, USA: Jossey-Bass, 1998.
[20] Z. Rohmah, H. Hamamah, E. Junining, A. Ilma, and L. A. Rochastuti, "Schools' support in the implementation of the Emancipated Curriculum in secondary schools in Indonesia," *Cogent Education*, vol. 11, no. 1, Art. no. 2300182, 2024, doi: 10.1080/2331186X.2023.2300182.
[21] T. Lickona, *Educating for Character: How Our Schools Can Teach Respect and Responsibility*. New York, NY, USA: Bantam, 1991.
[22] M. Kristjánsson, *Aristotle, Emotions, and Education*. Aldershot, U.K.: Ashgate, 2007.
[23] M. W. Berkowitz and M. C. Bier, "What works in character education," *Journal of Research in Character Education*, vol. 5, no. 1, pp. 29–48, 2007.
[24] Y. Anand *et al.*, "GPT4All: An ecosystem of open source compressed language models," in *Proc. 3rd Workshop on Natural Language Processing and Open Source Software (NLP-OSS)*, 2023, pp. 59–64, doi: 10.18653/v1/2023.nlposs-1.7.
[25] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *arXiv preprint arXiv:2005.11401*, 2020, doi: 10.48550/arXiv.2005.11401.
[26] UNESCO, *Guidance for Generative AI in Education and Research*. Paris, France: UNESCO, 2023, doi: 10.54675/EWZM9535.
[27] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang, "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education," *Smart Learning Environments*, vol. 10, Art. no. 15, 2023, doi: 10.1186/s40561-023-00237-x.
[28] OECD, *OECD Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem*. Paris, France: OECD Publishing, 2023, doi: 10.1787/c74f03de-en.
[29] N. Selwyn, *Should Robots Replace Teachers? AI and the Future of Education*. Cambridge, U.K.: Polity Press, 2019.
[30] W. Holmes, K. Porayska-Pomsta, K. Holstein, *et al.*, "Ethics of AI in education: Towards a community-wide framework," *International Journal of Artificial Intelligence in Education*, vol. 32, pp. 504–526, 2022, doi: 10.1007/s40593-021-00239-1.
[31] V. González-Calatayud, P. Prendes-Espinosa, and R. Roig-Vila, "Artificial intelligence for student assessment: A systematic review," *Applied Sciences*, vol. 11, no. 12, Art. no. 5467, 2021, doi: 10.3390/app11125467.
[32] H.-F. Hsieh and S. E. Shannon, "Three approaches to qualitative content analysis," *Qualitative Health Research*, vol. 15, no. 9, pp. 1277–1288, 2005, doi: 10.1177/1049732305276687.
[33] D. B. Yarbrough, L. M. Shulha, R. K. Hopson, and F. A. Caruthers, *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users*, 3rd ed. Thousand Oaks, CA, USA: Corwin Press, 2010.
[34] T. Langenfeld, "Internet-based proctored assessment: Security and fairness issues," *Educational Measurement: Issues and Practice*, vol. 39, no. 3, pp. 24–27, 2020, doi: 10.1111/emip.12359.
[35] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao, "Conceptualizing AI literacy: An exploratory review," *Computers and Education: Artificial Intelligence*, vol. 2, Art. no. 100041, 2021, doi: 10.1016/j.caeai.2021.100041.
[36] H. Crompton and D. Burke, "Artificial intelligence in higher education: The state of the field," *International Journal of Educational Technology in Higher Education*, vol. 20, Art. no. 22, 2023, doi: 10.1186/s41239-023-00392-8.
[37] R. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT," *Innovations in Education and Teaching International*, vol. 61, no. 2, pp. 228–239, 2024, doi: 10.1080/14703297.2023.2190148.
[38] J. Roe, M. Perkins, and Y. Tregubova, "The EAP-AIAS: Adapting the AI Assessment Scale for English for Academic Purposes," *TESOL Journal*, vol. 17, no. 2, Art. no. e70122, 2026, doi: 10.1002/tesj.70122.
[39] M. Q. Patton, *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*, 4th ed. Thousand Oaks, CA, USA: SAGE, 2015.
[40] T. Nazaretsky, M. Ariely, M. Cukurova, and G. Alexandron, "Teachers' trust in AI-powered educational technology and a professional development program to improve it," *British Journal of Educational Technology*, vol. 53, no. 4, pp. 914–931, 2022, doi: 10.1111/bjet.13232.
[41] M. Mangal and Z. A. Pardos, "Implementing equitable and intersectionality-aware ML in education: A practical guide," *British Journal of Educational Technology*, vol. 55, no. 5, pp. 2003–2038, 2024, doi: 10.1111/bjet.13484.
[42] A. Barrett and A. Pack, "Not quite eye to A.I.: Student and teacher perspectives on the use of generative artificial intelligence in the writing process," *International Journal of Educational Technology in Higher Education*, vol. 20, Art. no. 59, 2023, doi: 10.1186/s41239-023-00427-0.

APPENDICES

Appendix A. Formative Routine Checklist (FRC)
Purpose: Strengthen consistency of formative assessment cycles in low-connectivity classrooms.
Intended use: Teacher self-use; peer coaching; lesson planning reflection.
Mode: Paper-first (offline); optionally digitised.
Frequency: 1–3 times per lesson or at a minimum once per week per class.
Lesson/Topic: **Date:** / / **Class:** **Teacher:**
A1. Clarify the learning intention

- ☐ Learning intention is stated in student-friendly language
- ☐ Success criteria are explicit (what "good" looks like)
- ☐ Criteria are feasible given time/resources

A2. Elicit evidence (select at least one)

- ☐ Hinge question (whole-class)
- ☐ Exit ticket (end-of-lesson)
- ☐ Mini whiteboard / show-of-hands response
- ☐ Short written response (2–5 minutes)
- ☐ Observation checklist during task
- ☐ Peer check using criteria

Prompt/task used:
A3. Interpret evidence

- ☐ Evidence was checked during the lesson (not only after)
- ☐ Misconceptions/errors were identified
- ☐ Students needing immediate support were identified

What did the evidence show? (2–3 bullets)

-
-
-

A4. Take responsive action (must complete)

- ☐ Whole-class clarification / worked example
- ☐ Targeted small-group support
- ☐ Additional practice task
- ☐ Extension task for ready learners
- ☐ Other:

Responsive action taken (write exactly what changed):
A5. Close the loop

- ☐ Students were told what to do next
- ☐ Students acted on feedback (revision/resubmission/second attempt)
- ☐ Next lesson plan updated based on evidence

Next step for students (one sentence):
A6. Quick ratings (circle)
Evidence sufficiency: 0 1 2 3 (0 none; 3 strong and aligned)
Responsiveness: 0 1 2 3 (0 no change; 3 targeted and documented)
Equity check: ☐ Yes ☐ No (All learners had an accessible way to show evidence)
If no, barrier:
Validity note (brief): Intended for formative improvement (not grading). Primary threat is "feedback without action". Mitigation is a mandatory response-action documentation and evidence-first prompts.

Appendix B. Summative Task Quality Rubric (STQR)
Purpose: Improve summative task alignment, clarity, feasibility, and integrity resilience in low-connectivity settings.
Intended use: Task design; moderation; programme QA.
Mode: Paper-first (offline); optionally digitised.
Task title: **Year level:** **Subject:**
Curriculum objective(s):
Task type: ☐ Test ☐ Performance ☐ Project ☐ Portfolio ☐ Other:
B1. Quality rubric (rate each criterion 1–4)
Scale: 1 Needs major revision | 2 Needs revision | 3 Good | 4 Strong

1. **Alignment to objectives** (task evidences stated objectives) 1 2 3 4 Notes:
2. **Cognitive demand and authenticity** (requires application/decision-making) 1 2 3 4 Notes:

Evidence sufficiency (outputs allow a valid judgement; evidence not under-specified) 1 2 3 4 Notes:

3. **Clarity of instructions and criteria** (students know what to do; criteria explicit) 1 2 3 4 Notes:
4. **Fairness and accessibility** (reasonable adjustments possible; no hidden disadvantage) 1 2 3 4 Notes:
5. **Feasibility in low-connectivity contexts** (materials, devices, time, printing is realistic) 1 2 3 4 Notes:

6. **Integrity resilience** (design reduces copying and unauthorised assistance) 1 2 3 4 Notes:
7. **Marking reliability** (rubric supports consistent marking across markers) 1 2 3 4 Notes:

B2. Integrity design prompts (complete)

- **Process evidence requirement** (drafts, logs, checkpoints):
- **Contextual decision requirement** (localised choices hard to copy):
- **Verification point** (oral check, demonstration, conference, in-class task):

B3. Moderation record

Moderator: **Date:** / /

Required revisions (if any):

Validity note (brief): Intended for task quality and moderation (not student grading). Primary threats are construct under-representation and integrity vulnerability. Mitigation includes process evidence, verification points, and moderation.

Appendix C. Assessment Literacy Self-Audit (ALSA)

Purpose: Scaffold educator assessment literacy (criteria clarity, evidence sufficiency, fairness checks, responsive use of evidence).

Intended use: Individual reflection; peer coaching; PD tool.

Mode: Paper-first (offline); optionally digitised.

Assessment title: **Context:** ☐ Formative ☐ Summative ☐ Both

Date: / / **Teacher:** **Year level:**

C1. Self-audit (rate 0–3)

Scale: 0 Not yet | 1 Partly | 2 Mostly | 3 Consistently

1. I can state what learning this assessment is designed to evidence. 0 1 2 3
2. My criteria describe observable features of quality, not vague traits. 0 1 2 3
3. I can explain what evidence would change my judgement. 0 1 2 3
4. I anticipate common misconceptions and plan how to respond. 0 1 2 3
5. I can justify that the task is fair and accessible to diverse learners. 0 1 2 3
6. I separate achievement evidence from behaviour/compliance. 0 1 2 3
7. I provide actionable feedback (what to do next). 0 1 2 3
8. I use evidence to adjust teaching (not only to grade). 0 1 2 3
9. I can describe integrity risks and how design reduces them. 0 1 2 3
10. I can apply the rubric consistently and explain my decisions. 0 1 2 3

C2. Short reflection (complete all)

- **Strongest aspect of my current assessment design:**
- **One specific change I will make next time:**
- **Support I need (resources, exemplars, moderation):**

Validity note (brief): Intended to improve teacher judgement and design (not to evaluate teacher performance). Primary threat is superficial completion. Mitigation is specific action planning and peer coaching discussion.

Appendix D. Character Education Construct Map + Rubric (CECMR)

Purpose: Support credible, culturally grounded, multi-source assessment of character education outcomes.

Intended use: Primarily formative; summative use only with moderation and explicit evidence rules.

Mode: Paper-first (offline); optionally digitised.

D1. Locally endorsed construct selection

Select up to **four** locally endorsed constructs (replace examples as needed).

Construct 1:
Construct 2:
Construct 3:
Construct 4:

D2. Evidence rule (must meet)

For each construct, use **at least 2 evidence sources** across **at least 2 occasions**.

Evidence sources (tick those used):

- ☐ Student reflection (structured prompt)
- ☐ Peer feedback (structured criteria)
- ☐ Teacher observation notes (structured)
- ☐ Product artefact (work sample)
- ☐ Process evidence (drafts, roles, logs)
- ☐ Community/parent input (if appropriate and consented)

D3. Rubric (complete per construct)

Construct: **Period:**

Scale: 1 Emerging | 2 Developing | 3 Demonstrating | 4 Sustaining

1. **Understanding** (explains expectation/value in locally meaningful terms) 1 2 3 4 Evidence notes:
2. **Application in context** (acts consistently with construct in relevant situations) 1 2 3 4 Evidence notes:
3. **Reflection and growth** (acts on feedback; identifies improvement steps) 1 2 3 4 Evidence notes:
4. **Respect and cultural fit** (aligns with agreed norms; avoids harm) 1 2 3 4 Evidence notes:

D4. Anti-gaming safeguards (tick all applied)

- ☐ Reflection prompts require specific examples (who/what/when)
- ☐ At least one evidence source is teacher-observed
- ☐ Evidence includes process (not only a polished product)
- ☐ Peer feedback uses clear criteria (not popularity)
- ☐ Student can respond to evidence (procedural fairness)

Validity note (brief): Intended for developmental feedback (not ranking). Primary threats are cultural bias and performativity. Mitigation is community endorsement, multi-source rules, and moderation.

Appendix E. Deeper Learning Portfolio Template (DLPT)
Purpose: Assess deeper learning and student development over time using offline-feasible evidence, with integrity protections.
Intended use: Portfolio across a unit/term/semester.
Mode: Paper-first (offline); optionally digitised.
Student: **Class:** **Teacher:**
Portfolio period: / / to / /
E1. Portfolio entries (minimum three)
For each entry, include **process documentation**.
Entry title: **Date:** / /
Competency focus (tick):

- ☐ Problem-solving ☐ Reasoning ☐ Communication ☐ Collaboration ☐ Creativity ☐ Self-management

1. **Task and goal (2–3 sentences):**
2. **Evidence included (tick):**

- ☐ Draft 1 ☐ Draft 2 ☐ Final
- ☐ Planning notes/outline
- ☐ Feedback received (teacher/peer)
- ☐ Revision notes (what changed and why)
- ☐ In-class checkpoint artefact
- ☐ Oral verification record (conference notes)

3. **Reflection (answer all):**

- What decision improved the work? Why?
- What feedback did you use? What changed?
- What would you do differently next time?

E2. Teacher verification (required for integrity)
Verification method (tick): ☐ Oral check ☐ In-class demonstration ☐ Conference ☐ Other:
Date: / /
Teacher notes (brief):
E3. Portfolio rubric (rate 1–4)
Scale: 1 Emerging | 2 Developing | 3 Proficient | 4 Advanced

1. Quality of reasoning/decisions: 1 2 3 4
2. Use of evidence and feedback: 1 2 3 4
3. Process transparency (drafts/checkpoints): 1 2 3 4
4. Communication and clarity: 1 2 3 4
5. Reflection specificity and actionability: 1 2 3 4

Overall judgement: 1 2 3 4 **Comments:**
Validity note (brief): Intended to evidence growth over time. Primary threat is ghost-writing or unauthorised assistance. Mitigation is required process evidence plus teacher verification.

Appendix F. Minimum Viable Governance Checklist (MVGC)
Purpose: Operational checklist for the minimum viable governance package required for policy-defensible offline AI use in an offline OER programme.
Intended use: Programme readiness and ongoing audit.
Mode: Paper-first (offline); optionally digitised.
F1. Governance controls (tick when implemented)
Prompt library governance

- ☐ Approved prompt library exists and is versioned
- ☐ Prompts aligned to local curriculum and assessment norms
- ☐ Evidence-first prompts prioritised for assessment-related tasks

Role-based access

- ☐ Teacher/student/admin roles defined
- ☐ Student access restricted for summative assessment contexts
- ☐ Acceptable-use rules provided to staff and students

Retrieval scope restriction

- ☐ Retrieval constrained to the local OER/training corpus

- ☐ Corpus inclusion/exclusion criteria documented
- ☐ Open-ended generation restricted/disabled for designated assessment uses

Quality assurance checks

- ☐ Locally curated exemplars and rubrics reviewed periodically
- ☐ Bias/cultural mismatch checks conducted on constructs and prompts
- ☐ Error reporting and correction process in place

Update protocol

- ☐ Model/corpus update schedule documented
- ☐ Rollback capability available
- ☐ Change log maintained (what changed, why, who approved)

Integrity protocols

- ☐ Assessment designs include process evidence requirements
- ☐ Verification points planned (oral checks, conferences, demonstrations)
- ☐ Moderation routines documented for higher-stakes tasks

Accountability documentation

- ☐ Decision trail recorded (approvals, revisions, governance reviews)

☐ Periodic internal review scheduled (termly or semesterly)