E-ISSN: 2964-6227, P-ISSN: 2964-805X
Volume 3 Number 2 July 2025.
pp: 78-98. DOI: https://doi.org/10.58477/dj.v3i2.320.

Design Journal

RESEARCH ARTICLE Open Access

A Hybrid "Probabilistic Scaffold" Model for Enhancing Reading Literacy, Numeracy, and AI Literacy among Indonesian Primary School Students

Yusmaneli [1], **Fathurrahmad** [2*], **Ayu Raihani** [3].
[1] Junior Researcher, Research and Publication Division, Lembaga Mitra Solusi Teknologi Informasi (LMSTI), Banda Aceh City, Aceh Province, Indonesia.
[2*] Informatics Management Study Program, STMIK Indonesia Banda Aceh, Banda Aceh City, Aceh Province, Indonesia.
[3] Physics Education Study Program, Faculty of Teacher Training and Education, Universitas Syiah Kuala, Banda Aceh City, Aceh Province, Indonesia.

*Correspondence email: fathurrahmad@stmiki.ac.id

Received: 28 June 2025
Accepted: 25 July 2025
Published: 31 July 2025

Full list of author information is available at the end of the article.

Abstract
This design study develops a structured pedagogical framework positioning generative AI as a verification-dependent learning support for Indonesian 4th-5th grade students. The framework implements a sequential process (Prompt → Verify → Diagnose → Revise → Reflect → Transfer) bounded by ethical guidelines and local cultural contexts. The research addresses student bypassing of initial reasoning, absence of age-appropriate AI literacy assessment, digital access disparities, insufficient verification practices, and potential erosion of academic integrity. Using a quasi-experimental design (n≈200 across 4 schools; 12 weeks), the study compares an AI-integrated curriculum with conventional instruction, controlling for pre-test performance, socioeconomic indicators, and gender. The intervention balances AI-mediated learning (25-35%) with analog activities (65-75%) to preserve direct cognitive engagement. Assessment instruments include reading comprehension measures, contextual mathematics problems, a four-dimensional AI literacy rubric, interaction logs, classroom observations, motivation scales, and verification records. Anticipated outcomes include normalized learning gains of 0.35-0.45 in reading and 0.30-0.40 in mathematics; advancement to Level 3 AI literacy; reduction in AI dependence from 60% to 35-40%; increased higher-order questioning; and equitable outcomes across settings through offline resource distribution. The theoretical framework proposes that prompt quality influences learning outcomes through verification practices, metacognitive awareness, and ethical understanding. Implementation challenges include novelty effects, variable teacher adherence, technology access inconsistencies, and ethical standards application. The research aims to shift elementary AI education from answer retrieval toward evidence-based inquiry with ethical awareness.

Keywords: AI Literacy Assessment; Verification Protocols; Contextual Mathematics; Metacognitive Development; Educational Ethics.

Abstrak
Penelitian ini mengembangkan kerangka pedagogis terstruktur yang memposisikan AI generatif sebagai pendukung pembelajaran berbasis verifikasi untuk siswa kelas 4-5 SD di Indonesia. Kerangka ini menerapkan proses berurutan (Prompt → Verifikasi → Diagnosis → Revisi → Refleksi → Transfer) yang dibatasi oleh pedoman etika dan konteks budaya lokal. Penelitian menangani kecenderungan siswa melewati penalaran awal, ketiadaan penilaian literasi AI sesuai usia, kesenjangan akses digital, praktik verifikasi tidak memadai, dan potensi erosi integritas akademik. Menggunakan desain kuasi-eksperimental (n≈200 di 4 sekolah; 12 minggu), studi membandingkan kurikulum terintegrasi AI dengan pengajaran konvensional, mengendalikan

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

kinerja pra-tes, indikator sosial-ekonomi, dan gender. Intervensi menyeimbangkan pembelajaran berbantuan AI (25-35%) dengan kegiatan analog (65-75%) untuk mempertahankan keterlibatan kognitif langsung. Instrumen penilaian mencakup pengukuran pemahaman membaca, masalah matematika kontekstual, rubrik literasi AI empat dimensi, catatan interaksi, observasi kelas, skala motivasi, dan catatan verifikasi. Hasil yang diantisipasi meliputi peningkatan pembelajaran ternormalisasi 0,35-0,45 dalam membaca dan 0,30-0,40 dalam matematika; kemajuan ke Literasi AI Level 3; pengurangan ketergantungan AI dari 60% menjadi 35-40%; peningkatan pertanyaan tingkat tinggi; dan hasil yang setara antar lingkungan melalui distribusi sumber daya offline. Kerangka teoretis mengusulkan bahwa kualitas prompt memengaruhi hasil pembelajaran melalui praktik verifikasi, kesadaran metakognitif, dan pemahaman etis. Tantangan implementasi meliputi efek kebaruan, kepatuhan guru bervariasi, inkonsistensi akses teknologi, dan penerapan standar etika. Penelitian bertujuan mengalihkan pendidikan AI dasar dari pengambilan jawaban menuju penyelidikan berbasis bukti dengan kesadaran etis.

Kata Kunci: Penilaian Literasi AI; Protokol Verifikasi; Matematika Kontekstual; Pengembangan Metakognitif; Etika Pendidikan.

1. Introduction

Rapid advances in generative artificial intelligence, seen in emerging large language models like prospective GPT-5, are transforming what primary-age students can accomplish. Yet hasty adoption without structured pedagogical boundaries risks shallow language imitation, diminished mathematical effort, and subtle replacement of teacher judgment with opaque algorithmic guidance. Indonesian elementary education faces the challenge of embedding AI literacy as a core habit—viewing probabilistic text generation as adjustable support rather than an authoritative source—while bolstering reading fluency, mathematical adaptability, conversational exchange, ethical judgment, empathy, and determination aligned with Merdeka Belajar reforms and the Pancasila Student Profile, which values critical thinking, responsible creativity, mutual collaboration, cultural diversity appreciation, and self-direction. Converting these policy goals into classroom practice requires thoughtful design across curriculum planning, graduated skill development, balanced task environments (digital versus hands-on, verbal, experiential), redefined teacher roles, and risk management. Current obstacles include varying teacher capacity to create effective prompts, monitor semantic changes, and demonstrate verification methods (Karataş *et al.*, 2024; Alshehri, 2023); uneven infrastructure causing unpredictable technology access (Anggreni *et al.*, 2023); lack of child-specific privacy standards beyond basic consent forms (Weidener & Fischer, 2023); missing detailed AI literacy standards for upper elementary grades (Biagini, 2024); and growing dependency patterns where students skip initial reasoning and accept first-round AI outputs (Ali & Mughari, 2024).

Global child-focused AI literacy programs emphasize demystification—explaining systems as statistical word predictors prone to errors—combined with prompt creation, repeated revision, fact verification, and basic ethical understanding (Brummelen *et al.*, 2021; Druga *et al.*, 2019). Meanwhile, surveys across healthcare training programs highlight growing professional expectations to both use and evaluate AI outputs (Al Qerem *et al.*, 2023; Lee *et al.*, 2021; Moldt *et al.*, 2023; Kim *et al.*, 2025). However, directly applying advanced competency frameworks would overwhelm 9-12 year olds unless simplified into observable actions: recognize (distinguish AI from human writing), use (apply tools with guidance for vocabulary expansion or problem rephrasing), check (compare results against textbooks, experiments, or calculations), and reflect (describe limitations, bias risks, ethical boundaries) (Lee *et al.*, 2023; Kim *et al.*, 2025). Any blended approach must connect AI-assisted activities to locally meaningful themes such as school water conservation, market mathematics, regional storytelling adaptation, or native plant investigation to maintain relevance and richness (Sukmawati & Wahjusaputri, 2024; Anggreni *et al.*, 2023). Character development research suggests that moral formation in digital environments depends on clear standards for citation honesty, respectful communication, responsible paraphrasing, avoiding stereotype perpetuation, and data caution (Damanhuri & Wali, 2024).

The adaptive possibilities of large language models—automatically leveled reading passages, varied math problems within skill ranges, responsive vocabulary assistance—can reduce repetitive differentiation workload (Chan & Zary, 2019; Kim *et al.*, 2025) only when teachers possess evaluative skills to assess cognitive challenge, cultural appropriateness, and logical sequencing, abilities developed through structured learning communities and reflective coaching partnerships (Alshehri, 2023; Nurbaya *et al.*, 2023). Without clear boundaries, superficial usage patterns may emerge: quick summarization replacing careful reading, AI-generated narratives supplanting student drafting practice, direct answer requests

hindering numerical understanding development. To prevent such tendencies, we propose guiding principles: requiring initial independent thinking before AI assistance (Ali & Mughari, 2024), limiting AI-mediated activities (25-35% of weekly academic time), creating annotated verification records where students mark fact-checked claims, holding regular "hallucination analysis" sessions examining flawed outputs to normalize skepticism, maintaining structured prompt journals tracking progression from vague to precise questioning, and implementing bias review checklists examining portrayal of gender roles, occupations, and cultural aspects (Rainey *et al.*, 2021). Thorough assessment requires quasi-experimental comparison between AI-integrated instruction and conventional teaching to measure differences in reading comprehension accuracy, mathematical procedural fluency, higher-level questioning frequency, and self-efficacy changes while accounting for novelty effects (Park & Kwon, 2023). Assessment innovations adapt multi-dimensional AI literacy frameworks (Biagini, 2024) into child-appropriate rubrics covering conceptual understanding (*e.g.*, "AI predicts words" level), practical application (appropriate prompt specificity), evaluative skill (identifying omissions, incorrect facts), and ethical awareness (avoiding personal information sharing, acknowledging AI assistance).

Teacher professional identity evolves from primary feedback provider on simple tasks to orchestrator of learning apprenticeships where comparing student solutions with AI-generated guidance stimulates metacognitive monitoring and error detection. Safety measures must progress beyond generic disclaimers toward concrete protocols: anonymized user identification, approved prompt categories prohibiting personal address entry, tracking of rejection instances to refine policies, curated datasets for offline or limited-bandwidth schools, and regular external privacy reviews (Weidener & Fischer, 2023; Rainey *et al.*, 2021). Equity demands parallel low-technology options—printed materials matching adaptive levels, teacher-selected "frozen" AI outputs reviewed beforehand, scheduled device rotation maintaining collaborative discussion—to prevent educational divides (Anggreni *et al.*, 2023). Through this problem analysis, our research focuses on questions about literacy and numeracy improvement under AI scaffolding, shifts toward higher-order questioning (analysis, evaluation, creation), teacher perceptions of pedagogical control and role adequacy, effectiveness of privacy and bias mitigation protocols, motivation and self-efficacy trajectories, and varying impacts across geographic or resource differences. The conceptual framework suggests adaptive AI scaffolds trigger student reasoning attempts, verification processes refine mental models, character education shapes ethical tool use attitudes, and teacher guidance maintains intellectual vigilance, leading toward measurable gains in foundational skills and emerging AI literacy.

Table 1. Problem Space and Strategic Response

Challenge Category	Specific Issue	Strategic Response	Pedagogic Rationale	Key References
Teacher Readiness	Limited prompt engineering & output auditing skill	Targeted microcredential modules + peer coaching cycles	Builds diagnostic capacity & shared practice language	Karataş *et al.*, 2024; Alshehri, 2023
Learner Overreliance	Immediate answer seeking	Pre-AI attempt rule + reflective prompt logs	Preserves productive struggle & metacognition	Ali & Mughari, 2024
Bias & Representation	Stereotyped narratives	Bias checklist + periodic hallucination clinics	Normalizes critical inspection & inclusive framing	Rainey *et al.*, 2021
Privacy & Safety	Unstructured data sharing	Data minimization, whitelisted prompts, anonymized logs	Protects minors & builds trust	Weidener & Fischer, 2023
Assessment Gaps	No AI literacy rubric for ages 9–12	Adapt four dimension rubric (Concept, Use, Check, Ethics)	Enables formative feedback & progression tracking	Biagini, 2024
Infrastructure Inequity	Device scarcity, low bandwidth	Rotational use + offline vetted output packs	Ensures equitable exposure	Anggreni *et al.*, 2023
Motivation & Self Efficacy	Possible passivity from automation	Reflection sheets + success attribution coaching	Reinforces agency & resilience	Park & Kwon, 2023
Cultural	Disconnect from	Thematic tasks	Anchors abstraction	Sukmawati &

Relevance	local themes	(markets, water, flora) + AI variation	in lived experience	Wahjusaputri, 2024
Character Formation	Untracked ethical practice	Explicit norms (attribution, respectful tone) + rubric	Integrates moral development with tool use	Damanhuri & Wali, 2024
Evaluation Rigor	Attribution of gains uncertain	Quasi experimental design (ANCOVA) + mixed methods	Isolates intervention signal	Park & Kwon, 2023

Table 2. Proposed AI Literacy Progression (Simplified)

Dimension	Grades 1–2	Grades 3–4	Grades 5–6	Observable Indicators
Recognize	Distinguish tool vs human	State AI predicts text patterns	Explain probability & possible errors	Verbal explanation accuracy
Use	Guided simple question	Formulate specific prompt (who, what, why)	Optimize prompt for clarity & constraint	Prompt quality rubric score
Check	Ask adult "Is this right?"	Compare output to textbook / experiment	Triangulate across 2+ sources & annotate discrepancies	Verification log frequency
Reflect / Ethics	Avoid sharing name	Identify unsafe personal info	Articulate bias risk & cite AI assistance	Ethical rubric level

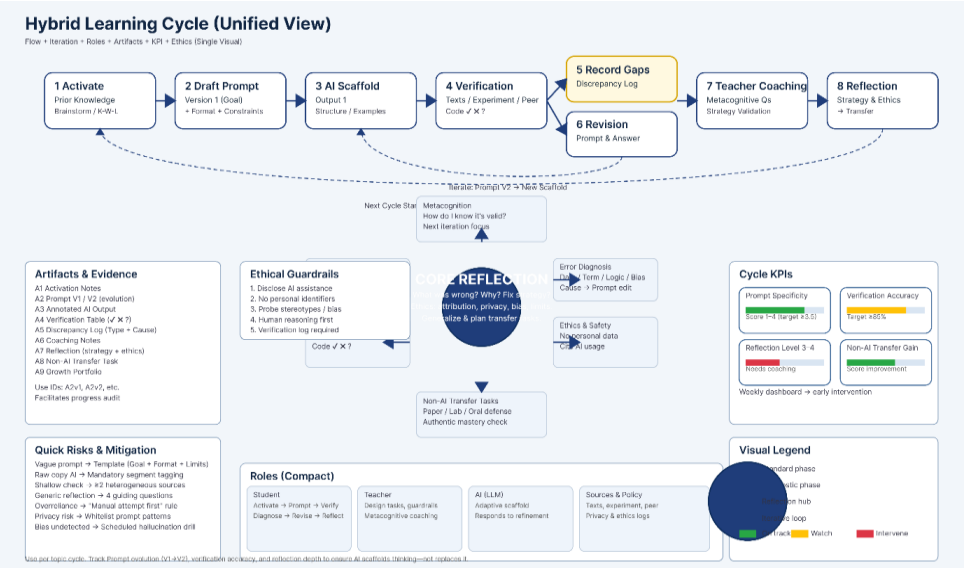

Figure 1. Hybrid Learning Cycle – Unified Operational Map

Figure 1 integrates five crucial layers: (1) A sequential content flow (Activate → Draft Prompt → AI Scaffold → Verification → Record Gaps → Revision → Teacher Coaching → Reflection) tracking knowledge progression from initial assumptions to strategy consolidation; (2) A cyclical process asserting that prompt and output quality improves through evidence-based diagnosis-revision, not random attempts; (3) A central Reflection Hub for metacognitive regulation where students examine errors, causes, improvements, and strategy transfer; (4) Supporting frameworks (Artifacts & Evidence, Roles, KPIs, Risk & Mitigation, Ethical Guardrails) forming an accountability system—each artifact coded (A1–A9) for longitudinal tracking, with KPIs serving as formative indicators for timely teacher intervention; (5) Ethics & Security foundations for AI integration while preserving academic integrity, privacy, and bias awareness. The resulting operational structure balances creative discovery, verification discipline, strategy refinement, and independent mastery (non-AI transfer tasks). This framework serves as a classroom guide, teacher orientation resource, or evaluation tool for school AI literacy programs. Our study develops a purposefully crafted hybrid model where generative AI functions as adaptable support rather than an answer provider;

verification and reflection build intellectual vigilance; ethical guidelines preserve integrity; and teacher coaching refocuses human educational judgment—all targeting measurable improvements in basic literacy, mathematical flexibility, and emerging AI literacy for upper elementary students in Indonesia. Subsequent sections detail theoretical foundations, model design, and methodological approaches examining these propositions.

2. Methodology

2.1 Research Design

This study employs a multi-group quasi-experimental design with a 12-week core intervention period to evaluate the effects of a Hybrid AI Curriculum on reading literacy, numeracy, AI literacy, learning motivation, and character indicators. Two conditions are compared: (a) Experimental group receives the Hybrid AI Curriculum integrating thematic modules and a cognitive cycle (Prompt → Verification → Diagnosis → Revision → Reflection → Transfer) within a curated GPT-5 sandbox; (b) Control group follows the conventional curriculum without generative AI support but with equivalent instructional time. Full randomization across classes isn't feasible due to operational school constraints; thus analyses will statistically control pre-test scores, simple SES indicator, and gender, and will apply multilevel (hierarchical) modeling when intraclass correlation coefficients (ICC) indicate meaningful clustering (class/school). Fidelity monitoring supports valid attribution of outcome differences to treatment differentiation. (Figure 1. Quasi-Experimental Design Scheme).

2.2 Setting and Participants

Four primary schools (2 urban with stable connectivity; 2 semi-rural with fluctuating connectivity) at Grades 4 and 5 are targeted (approximate total n ≈ 200 students). Purposive sampling criteria: willingness to participate, leadership support, minimal infrastructure (device ratio ≤ 1:4), and readiness to apply data ethics protocols. Baseline characteristics (gender, simple SES indicators—*e.g.*, device ownership, parental education, initial literacy/numeracy scores) are collected as covariates and to inspect potential selection bias. Inclusion criteria: active enrollment, baseline attendance ≥ 80%. Exclusion criteria: cumulative absence > 30% or intensive special needs requiring individualized adaptation (recorded descriptively, excluded from inferential analysis if n insufficient). Sample size rationale assumes a moderate effect (d ≈ 0.40); design effect due to clustering is acknowledged; confidence intervals will accompany effect estimates to temper overclaim.

2.3 Intervention

The Hybrid AI Curriculum balances 65–75% analogue activities (oral discussion, simple science experiments, market simulations, narrative drama, environmental observation) and 25–35% structured AI-supported activities (generating vocabulary variants, adaptive numerical scenarios, multi-voice narrative perspectives, scaffolding higher-order questioning). AI is positioned as a co-facilitator to stimulate elaboration—not as a final answer source. The GPT-5 sandbox enforces a domain whitelist (literacy, contextual numeracy, environment, local culture) and blocks (blacklists) personal data or sensitive content. Four thematic modules—(1) Regenerative School Garden; (2) Healthy Traditional Market; (3) Clean & Efficient Water; (4) Multi-Voice Nusantara Stories—follow a recurring sequence: activation → analogue exploration → AI interaction → verification → application/product → reflective transfer. Dual guardrails: (a) Pre-Reasoning (students articulate initial hypotheses/solution steps before AI access); (b) Verification Table (classifying AI claims: correct, incorrect, uncertain + source). These align with critical AI literacy and epistemic vigilance principles (Lee *et al.*, 2021; Brummelen *et al.*, 2021). Fidelity indices: teacher checklists (proportion of sessions implementing both guardrails), system logs (prompt revision ratios), structured observations. (Figure 2. Hybrid Learning Cycle Flow).

2.4 Teacher Professional Development

Capacity building spans three phases: (1) Orientation to AI concepts and limitations (hallucination risk, representational bias) grounded in educational AI literacy frameworks (Lee *et al.*, 2021; Brummelen *et al.*, 2021); (2) Prompt design & curation plus output auditing to enhance meaningful interaction (Karataş *et al.*, 2024); (3) Ongoing pedagogical reflection and peer coaching emphasizing teacher evolution toward orchestration roles (Alshehri, 2023). An online community of practice enables artefact sharing (effective prompts, exemplary verification episodes). Teacher roles are reframed: prompt curator, metacognition facilitator, bias auditor, ethics steward. Competence progression is internally tracked (Awareness → Functional → Orchestrator → Reflective).

2.5 Instruments and Indicators

Instruments include: (1) Reading Literacy Test assessing literal, inferential, evaluative comprehension and efficiency (correct score per minute). (2) Numeracy Test (mixed operations, contextual problem solving involving market and water measurement, estimation). (3) Four-Dimension AI Literacy Rubric—Concepts, Use, Evaluation/Verification, Ethics—adapted from Biagini (2024), Levels 1 (Basic) to 4 (Transformative). (4) AI Log Analytics coding prompt type (factual, conceptual, meta-reflective), number of revisions prior to manual validation, specificity trajectory, presence of pre-reasoning, and verification status. (5) Motivation & Self-Efficacy Scale (Likert 1–5) targeting Cronbach α > 0.70. (6) Structured Classroom Observation capturing high-order questions (Analysis, Evaluation, Creation), verification initiative, collaboration. (7) Semi-Structured Teacher Interviews (adapted from Lee *et al.*, 2023) eliciting perceptions of pedagogical control, cognitive load, adaptive strategies. (8) Character Rubric (Collaboration, Responsibility, Empathy) adapted from Anggreni *et al.* (2023). (9) Verification Table documenting evaluation decisions (claim correctness + reference). Table 1 summarizes instruments.

Table 1. Summary of Instruments, Dimensions, Key Indicators, and Analyses

Instrument	Dimensions / Sub-Dimensions	Key Indicators (Examples)	Format / Scale	Main Analyses (Inferential / Other)
Reading Literacy Test	Literal; Inferential; Evaluative; Efficiency	Correct answers per domain; time; efficiency index (score/min)	30 items; timed (MCQ ± short constructed)	ANCOVA; item difficulty & discrimination; KR-20; effect size
Numeracy Test	Operations; Contextual Problem Solving; Estimation	Total score; subscale scores; Rasch person ability	25 contextual items	Rasch (infit/outfit); ANCOVA; N-Gain; effect size
AI Literacy Rubric (Biagini, 2024)	Concept; Use; Evaluation / Verification; Ethics	Level (1–4) per dimension; % reaching ≥ Level 3	4-level analytic rubric	Ordinal logistic / CLMM; Cohen's d (derived); Kappa (subset)
AI Log Analytics	Prompt Type; Revision Count; Verification Status; Specificity; Pre-reasoning	Distribution of prompt types; mean revisions; specificity level; proportion with pre-reasoning	Automated metadata log	Descriptive; sequence analysis; (optional) GLMM over time
Motivation & Self-Efficacy Scale	Intrinsic Motivation; Persistence; Self-efficacy	Subscale scores; total score	Likert 1–5	ANCOVA; Cronbach's α; (optional) EFA/CFA; effect size
Classroom Observation	Higher-order Questions (Analysis/Evaluation/Creation); Verification Initiative; Collaboration	Session frequency counts; proportion higher-order questions	Structured coding sheet	Poisson / NegBin MLM; proportion (logit) models
Teacher Interview (Lee *et al.*, 2023)	Role Reconfiguration; Cognitive Load; Adaptation Strategies	Thematic clusters; representative quotes	Semi-structured transcript	Thematic analysis (Braun & Clarke); triangulation
Character Rubric (Anggreni *et al.*, 2023)	Collaboration; Responsibility; Empathy	Level (1–4) per dimension; % level gains	4-level rubric with anchors	Kappa; Ordinal logistic / cumulative link model
Verification Table	Claim Accuracy; Source Triangulation; Uncertainty Tagging	Proportion claims verified; accuracy after verification	Claim–status–source table	Mediation (verification → outcomes); logistic regression
Prompt	Prompt Specificity Evolution;	Specificity level (1–	Weekly	Multilevel growth

Journal	Strategy Shifts	4); weekly growth trajectory	narrative entries + coded level	model (MLM); correlations with outcomes

2.6 Instrument Development and Adaptation

Drafts underwent expert review (literacy, numeracy, AI pedagogy), pilot testing (external sample), item analysis (difficulty p ideal 0.30–0.80; discrimination ≥ 0.20), and revision. AI literacy and character rubrics include anchored exemplars per level to reduce rater ambiguity. Adaptations preserved content validity while localizing contexts (traditional market, environmental stewardship). Prompt type classification uses semi-automated keyword/pattern scripts with 10% manual validation (precision target ≥ 0.85). Prompt specificity rubric: Level 1 generic; Level 2 added attribute; Level 3 contextual + criteria; Level 4 multi-criteria + evaluative purpose.

2.7 Procedure

Week 0: Parental consent; reading & numeracy pre-tests; baseline AI literacy task (explain "what is AI" + verify a simple AI claim); motivation & self-efficacy scale; demographic survey. Weeks 0–1: Teacher training Phases 1–2; sandbox configuration (whitelist & filters). Weeks 1–12: Thematic module implementation. Weekly structure: (1) Literacy session (vocabulary variation via AI → synonym verification); (2) Contextual numeracy (adaptive market scenario → manual calculation → AI suggestion comparison); (3) Science/social project (analogue experiment → AI-suggested variation → filtering); (4) Verification reflection (analysis of incorrect claims + error patterns); (5) Character & multi-voice narrative (AI prompts perspective shifts → manual clarification). Every AI interaction requires prior Pre-Reasoning and at least two Verification Table entries. Classroom observations (systematic sampling) occur Weeks 2, 5, 8, 11. Week 12: Post-tests (literacy, numeracy, AI literacy), motivation scale re-administration, teacher interviews, Prompt Journal collection. Logs exported daily with anonymization hashing. (Figure 2. Hybrid Learning Cycle; Figure 3. Phased Implementation Timeline).

2.8 Data Analysis

Quantitative: (a) ANCOVA with post-test scores as dependents, controlling for pre-test, SES proxy, gender; (b) Multilevel modeling (MLM) if ICC > 0.05 (Level 1 students; Level 2 classes; school added if variance significant); (c) Effect sizes (Cohen's d; Hedges g when unequal n) and N-Gain (Hake) to interpret relative improvement; (d) High-order question proportion analyzed via generalized linear model (logit) with total questions offset; (e) Exploratory mediation: Prompt Specificity → Verification Accuracy → Literacy outcomes (bootstrap 5,000 samples). Assumptions (normality, linearity, homoscedasticity) assessed prior to inference; robust or transformation strategies applied if violated. Missing data > 5% examined (MCAR vs MAR); Multiple Imputation used when MAR plausible. Sequence analysis traces transitions: Pre-Reasoning → Prompt Type → Revision Iterations → Verification Outcome. Qualitative: Thematic analysis (Braun & Clarke adaptation) with combined deductive (a priori: teacher role, learner agency, cognitive load) and inductive coding for emergent themes. Cross-source triangulation (logs, observations, interviews) enhances credibility.

2.9 Validity and Reliability

Content validity assured through expert review; internal consistency targeted (Cronbach α > 0.70; KR-20 for reading test). Numeracy calibration via Rasch (infit/outfit 0.7–1.3; misfitting items revised/removed). Inter-rater reliability for AI literacy and character rubrics targets Kappa ≥ 0.75 (substantial). Member checking (brief) on teacher interview summaries confirms interpretive accuracy. Qualitative audit trail (coding versions, analytic memos, theme merge decisions) is maintained for transparency. Social desirability bias mitigated by clarifying results do not affect summative grading. Consequential validity monitored to prevent "AI drill" patterns (observations confirm sustained conceptual exploration).

2.10 Ethics, Privacy, and Security

Parental written consent and student assent obtained; students receive a child-friendly explanation of AI's probabilistic nature ("AI predicts words; it can be wrong"). Personal data minimized (initials + code only); linkage file encrypted separately. Logs store metadata (prompt type, token length, verification status) without sensitive textual content. Whitelisted prompts block full names, addresses, out-of-scope domains. Bias mitigation includes cross-checking with national textbooks and explicit discourse on representation (Rainey *et al.*, 2021). Quarterly internal audits review data minimization and incident logs. Compliance with national child protection guidelines enforced through role-based access and encrypted storage.

2.11 Risk Mitigation Strategy

Overreliance addressed via "Independent Attempt First" (mandatory pre-AI reasoning) and initial time-gating AI access. Device disparity in semi-rural schools mitigated through printed modules and pre-generated (frozen) AI outputs still requiring manual verification. Teacher resistance reduced through peer coaching and quick-win micro-reports (improved verification accuracy). Content bias countered by representation checklists (gender, culture, occupation). Privacy upheld via hashed IDs and aggregated log exports. Table 2 maps risks and mitigations.

Table 2. Risk Mapping and Mitigation

Risk	Operational Definition	Impact	Risk Level	Core Mitigation	Monitoring Indicator	Target
Overreliance	AI usage without student pre-reasoning	Reduced cognitive elaboration	Medium	Require Pre-Reasoning form	Overreliance Index = % prompts without pre-reasoning	Reduction ≥25% from baseline
Content Bias	Unbalanced representation in AI content	Distortion of contextual relevance	Medium	Bias checklist during verification	Bias Prompt Rate = % prompts needing correction	<5%
Device Gap	Unstable access hindering verification	Inconsistent experience	High	Pre-generated outputs; low-bandwidth mode	Verification Gap = % difference between locations	<10% gap
Teacher Resistance	Adaptation burden reducing implementation	Reduced fidelity	Medium	Peer coaching; exemplar repository	Fidelity Rate = % sessions meeting core components	≥80%
Privacy	Unauthorized personal data input	Data exposure risk	Medium	Domain whitelist; filters; log audits	Privacy Incident Count	≤2 incidents total
Low Fidelity	Verification stage systematically skipped	Reduced intervention effect	Medium	Verification checklist + feedback	Verification Rate = % AI sessions with verification	≥70%
Reflection Fatigue	Declining quality of reflection content	Reduced metacognitive gains	Medium	Varied reflection formats	Reflection Quality Δ = mean score change/SD	Decline ≤0.3 SD

Table 2 maps key risks potentially weakening intervention process and outcome validity. The first focus is Overreliance: its indicator (proportion of interactions without pre-reasoning) was chosen because pre-reasoning steps serve as internal quality control for cognitive engagement; target reduction ≥ 25% signals strategy change, not random fluctuation. Content Bias risk appears because contextual bias may reduce learning relevance across backgrounds; threshold < 5% keeps bias residual, not systemic. Device Gap is addressed through redundancy (pre-generated outputs) to maintain equitable access to verification procedures; verification frequency gap < 10% demonstrates equalization. Teacher Resistance is positioned as implementation risk: fidelity (≥ 80%) signals that core components (pre-reasoning, verification, reflection) remain intact. Privacy is limited through preventive mechanisms (whitelist, audit) so incidents (≤ 2) can be managed without reducing stakeholder trust. Low Fidelity in verification threatens evaluative learning mechanisms; thus minimum rate ≥ 70% of sessions containing verification becomes prerequisite before scaling. Reflection Fatigue is quantitatively monitored through standardized rubric scores; decline > 0.3 SD indicates reflection quality erosion requiring intervention (format rotation). Overall, this table provides a measurable "early warning system" for adaptive corrective action during pilot and scaling phases.

Table 3. Success Criteria Matrix

Code	Domain/Outcome	Indicator	Operational Definition	Target/Success Criteria	Data Source	Timing
SC1	Literacy & Numeracy	N-Gain	(Post−Pre)/(Max−Pre) per student	Mean N-Gain ≥0.30	Literacy & numeracy tests	Pre & Post
SC2	Classroom Interaction	High-Order Question Increase	Δ proportion of high-order questions	≥+15 percentage points	Classroom observation	Baseline, Mid, Post
SC3	AI Literacy	Proportion Level ≥3	% of students at Level 3 or higher	≥70%	AI Literacy Rubric	Mid & Post
SC4	Psychosocial	Self-Efficacy Improvement	Δ self-efficacy score; effect size d	p<0.05 & d≥0.40	Motivation & efficacy scale	Pre & Post
SC5	Verification Practice	Manual Verification Inclusion	% of AI interactions with verification	≥40%	AI Log + Observation	Weekly
SC6	Overreliance Reduction	Overreliance Index Reduction	% reduction in Overreliance Index	≥25% from baseline	AI Log	Weekly
SC7	Assessment Reliability	Inter-rater Kappa	Cohen's/Weighted Kappa on samples	≥0.75 (substantial)	Rubrics	Each calibration
SC8	Ethics/Privacy	Privacy Incidents	Number of verified incidents	≤2 total	Log audit & reports	Ongoing
SC9	Prompt Quality	Prompt Specificity Gain	Mean improvement in prompt level	≥+1 level average	Prompt journal + log	Weekly
SC10	Mechanism	Verification Mediation	Intervention→Verification→Outcome	Significant indirect effect	Log + Tests	

Success criteria are built across three layers: (a) learning outcomes (1, 2, 3, 4), (b) AI process and practice quality (5, 6, 9, 10), and (c) ethical-methodological integrity (7, 8). N-Gain (≥ 0.30) represents moderate improvement commonly used in literacy/numeracy program evaluation. Increased proportion of high-order questions (+15 p.p.) validates pedagogical shift from mere recall toward analysis and evaluation. AI Literacy Level ≥ 3 (≥ 70% students) marks internalization of reflective-verificative practices. Self-efficacy requires not just statistical significance ($p < 0.05$) but also practical meaning ($d \geq 0.40$) to avoid "significance without substance." Verification process indicator (≥ 40%) serves as foundation for mediation pathway (criterion 10) testing whether verification practices bridge intervention influence on cognitive outcomes. Overreliance reduction (≥ 25%) and prompt specificity increase (+1 level) show enhanced metacognitive strategies in AI utilization. Assessment reliability (Kappa ≥ 0.75) ensures rubric interpretation isn't shifted by inter-rater inconsistency. Privacy incident limit (≤ 2) maintains ethical and social legitimacy of the intervention. Integratively, these criteria provide an evaluation framework balancing academic effectiveness, process depth, and ethical accountability.

Table 4. Phased Implementation Plan

Phase	Duration (Weeks)	Main Objectives	Key Activities	Output / Deliverables	Data Collected (Brief)	Decision / Gate (Go/No-Go)	Lead Role
1 Pre-Pilot	4	Module co-design; instrument adaptation; sandbox configuration.	Design workshop; rubric adaptation; technical log testing; initial teacher	Draft modules; final rubrics; log protocol; teacher	Content validity (Aiken's V); technical test notes.	≥80% expert agreement; stable system.	Research Team + Core Teachers

			training; content validation (expert review).	guide v1.			
2 Pilot (Feasibility)	6	Test procedure feasibility & refine rubrics/logs in 2 pilot schools.	Limited implementation; observation; teacher feedback collection; adjustment of rubrics & pre-reasoning forms.	Revised rubrics; feasibility report; initial parameters.	Process data; baseline overreliance; initial reliability	Fidelity ≥70%; complete logs ≥90%.	Field Researcher
3 Scaling	12	Full implementation in 4 schools; main quantitative & qualitative data collection.	Routine implementation; risk monitoring; periodic coaching; pre/post test collection; observation; teacher interviews (theoretical sampling).	Complete dataset; risk notes; transcripts; structured logs.	Pre/post tests; verification logs; observation; journals	Data loss <10%; high risks controlled.	School Coordinator + Researcher
4 Evaluation & Dissemination	(Analysis & Dissemination)	Final analysis; interpretation; dissemination and initial policy recommendations.	Statistical & mediation analysis; report preparation; brief national guide creation; open module repository; stakeholder presentation.	Final report; policy brief; open repository; anonymized dataset	Analysis results; effects; CI; data quality documentation	≥80% success criteria achieved or deviation explained	Analysis & Dissemination Team

The phased design enables structured learning-by-doing cycles. Phase 1 emphasizes system readiness: instrument adaptation, content validation, and technical logging configuration so the pilot phase isn't burdened with fundamental fixes. Phase 2 evaluates operational feasibility in two different contexts to test procedure robustness (*e.g.*, log completeness, teacher burden). Progression to Phase 3 is based on "gate" decisions (*e.g.*, fidelity ≥ 70%) to reduce the risk of scaling with unstable procedures. Phase 3 functions as the main inference phase: quantitative data (pre/post, logs, observations) and qualitative data (interviews) are collected under more representative conditions. Risk monitoring runs parallel so deviations are intercepted before disrupting dataset integrity. Phase 4 combines confirmatory statistical analysis (main effects, effect sizes, mediation) with practical synthesis (policy brief and open repository) to ensure transferability and transparency. This structure aligns with evidence-based education implementation principles: iterative, adaptive, and documented.

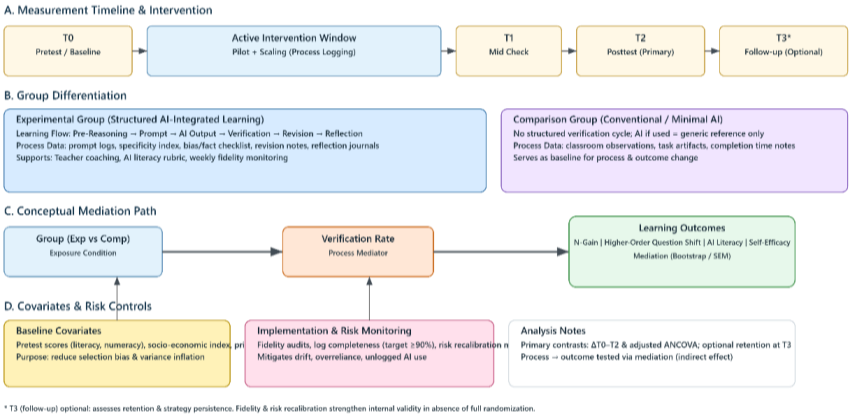

Figure 2. Quasi-Experimental Design Scheme

This quasi-experimental framework confronts the challenge of limited random assignment by intensifying process-level instrumentation. The experimental group is deliberately scaffolded through a structured cycle (Pre-Reasoning → Prompt → AI Output → Verification → Revision → Reflection), generating high-resolution logs that capture cognitive regulation, factual scrutiny, and iteration depth. The comparison group, lacking the enforced verification loop, establishes a counterfactual for both performance gains and process sophistication. Positioning verification rate as a mediator enables examination of mechanism: whether increased disciplined scrutiny mediates the relationship between structured AI integration and improved outcomes. Covariate adjustment (ANCOVA or SEM with observed covariates) statistically isolates treatment-related variance from pre-existing group differences. Fidelity metrics (*e.g.*, percentage of required verification checklists completed, prompt log completeness) serve dual roles—quality assurance and moderators in sensitivity analysis. Risk recalibration mid-implementation addresses potential drift (*e.g.*, unlogged AI queries, superficial verification) before it erodes internal validity. Optional follow-up (T3) interrogates durability: Do metacognitive and evaluative behaviors persist after the structured support window ends?. Analytically, mediation can be tested via bias-corrected bootstrap confidence intervals of indirect effects or latent variable modeling if measurement reliability thresholds (α or ω ≥ .70) are met. Should baseline imbalance remain after covariate control, propensity score weighting or overlap weighting may be appended as a robustness check. Reporting should transparently distinguish total effect, direct effect (controlling for verification rate), and indirect effect, clarifying the functional contribution of process quality to outcome attainment.

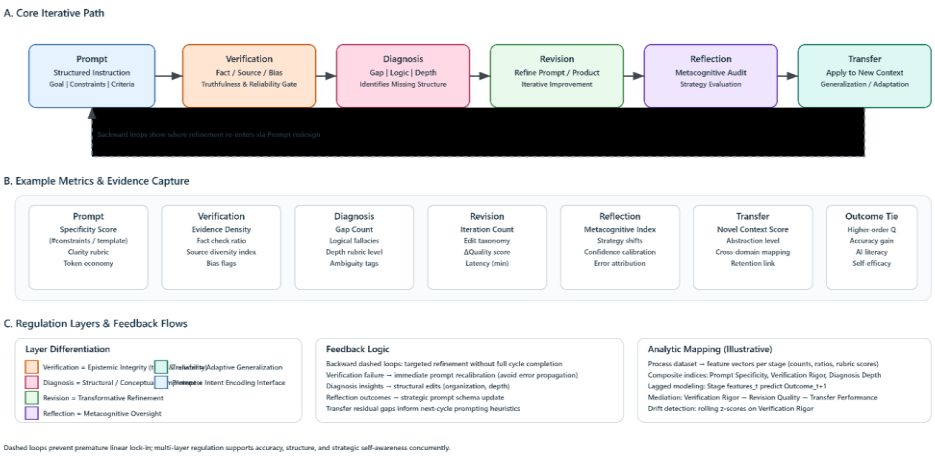

Figure 3. Hybrid Learning Cycle (Prompt → Verification → Diagnosis → Revision → Reflection → Transfer)

This cycle operationalizes a principled division of regulatory labor across five complementary layers: (1) Intent Encoding (Prompt), (2) Epistemic Integrity (Verification), (3) Structural Completeness (Diagnosis), (4) Transformative Refinement (Revision), (5) Metacognitive Oversight (Reflection), and (6) Adaptive Generalization (Transfer). The separation of Verification and Diagnosis prevents conflating factual fidelity with conceptual depth. Verification focuses on source grounding, factual accuracy, and bias detection; Diagnosis interrogates logical organization, conceptual gaps, and representational depth. Only after these gates pass does Revision

reconfigure content or the originating prompt, minimizing rework. Backward dashed loops instantiate selective recursion: failing a verification check produces prompt recalibration before users anchor on incorrect frames; structural deficiencies identified at Diagnosis trigger structural amendments without nullifying earlier epistemic checks; Reflection abstracts strategy adjustments (*e.g.*, adopting parameterized prompt templates) that feed forward into the next cycle. Transfer closes the loop by testing generalization—performance differentials across novel contexts reveal latent fragilities and inform subsequent prompt constraints or verification checklists. Analytically, each stage produces quantifiable features: Prompt Specificity (constraint density, clarity rubric), Verification Rigor (evidence density, source diversity, bias flags per 100 tokens), Diagnosis Depth (gap count normalized by expected concept map nodes), Revision Intensity (edit taxonomy diversity, Δquality), Reflection Metacognitive Index (strategy shift enumeration, confidence calibration error), and Transfer Performance (generalization ratio to baseline, abstraction level rubric). These can populate multilevel or structural equation models to: (a) test mediation chains (*e.g.*, Prompt Specificity → Verification Rigor → Transfer Performance), (b) detect temporal drift (rolling z-scores of verification rigor), and (c) evaluate diminishing returns (nonlinear saturation curves for iteration count vs quality gains). Pedagogically, the framework discourages over-reliance on AI outputs by interposing explicit verification before internalization, while Reflection and Transfer prevent inert knowledge by enforcing strategy articulation and contextual migration. The design is adaptable: low-resource contexts can collapse Diagnosis and Revision; advanced cohorts may parallelize Verification (AI-assisted fact checking) and Diagnosis (human conceptual review). Transparency of process logs supports reliability auditing, replication, and ethical oversight.

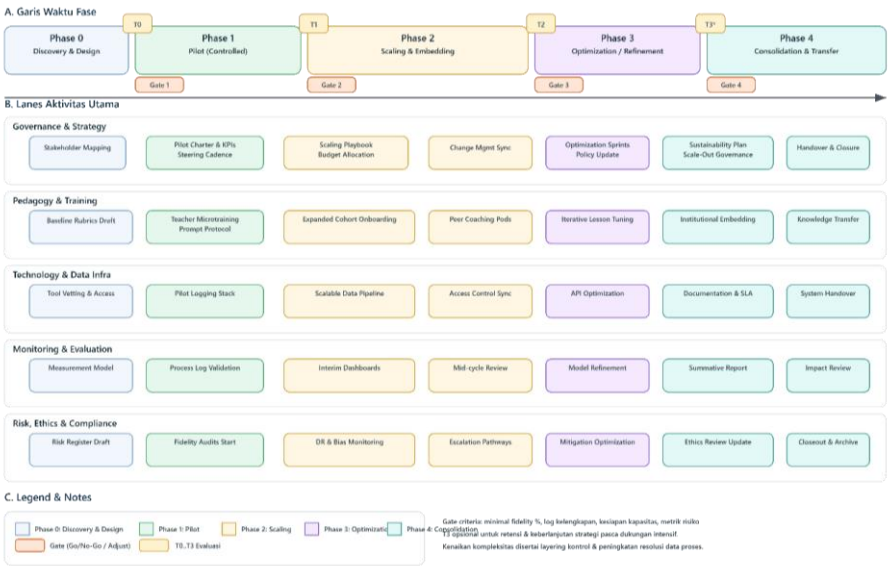

Figure 4. Phased Implementation Timeline

This phased design balances adoption speed with quality control through explicit Gate mechanisms. Phase 0 focuses on discovery (needs analysis, stakeholder mapping, basic rubric design) to build a cross-functionally aligned problem representation. Gate 1 examines minimal readiness: rubric validation, KPI definition, and basic logging readiness. Phase 1 (Pilot) emphasizes controlled experimentation: teacher microtraining, initial logging stack implementation, and initial fidelity audits to test process instrumentation (*e.g.*, log completeness >85%). Data at T0–T1 allows for initial impact estimation and design adjustments. Phase 2 expands the scope to more classes (scaling), requiring documented playbooks, more automated data pipelines, and integrated access control. The focus shifts from "does this work?" to "how can it be standardized without quality degradation?" Gate 2 evaluates process stability (fidelity variation across classes) and support capacity. Phase 3 focuses on optimization: improvement sprints, tuning the prompt protocol, and refining the analytical model (*e.g.*, increasing the granularity of verification and revision features). Gate 3 assesses whether marginal gains are still significant or have approached a plateau, warranting a transition to consolidation. Phase 4 shifts the emphasis to sustainability and transfer: final documentation, institutional embedding (internal policies, lesson plan/LMS integration), and a T3 evaluation that examines retention and persistence strategies (verification and diagnosis) without intensive support. Gate 4 closes the cycle with an ethics review and data archiving in accordance with privacy policies. The multi-lane structure emphasizes that success lies not solely in pedagogical activities, but in the synergy between layers: governance provides direction and allocation, training builds competency,

technology ensures data availability and quality, M&E translates logs into insights, while risk and ethics maintain integrity and compliance. The sequencing of activities within each lane minimizes the critical path—for example, scaling data pipelines before class volumes increase sharply. Analytically, the placement of evaluations T0–T2 allows for change modeling (ANCOVA/growth modeling), while T3 supports retention analysis. Gate metrics (fidelity percent, log completeness, interclass variance, bias incidence) can be treated as indicators of structural readiness and secondary predictors of cognitive outcomes or AI literacy. Phase transparency reduces the risk of scope creep and facilitates replication in other contexts.

3. Results and Discussion

3.1 Results

Since the project remains in its design phase, all values represent theoretically grounded projections informed by adjacent empirical and conceptual literature. First, the experimental group is expected to achieve a moderate normalized gain in reading literacy (N-Gain ≈ 0.35–0.45), attributable to pairing AI-assisted vocabulary differentiation with explicit verification routines that promote close reading instead of superficial compression. Second, numeracy performance is projected to rise with a moderate N-Gain (≈ 0.30–0.40), particularly on multi-step contextual mixed operations framed through adaptive traditional market price scenarios while preserving a compulsory manual reasoning pathway to prevent cognitive substitution. Third, student AI literacy is anticipated to shift from an average baseline at Levels 1–2 to Level 3 ("Progressive") on the Concept and Verification dimensions, operationalized as the ability to articulate that the model "predicts the next token probabilistically" and to independently locate at least two corroborating external sources for factual claims. Fourth, the proportion of higher-order questions (analysis, generalization, evaluative comparison, or clarification probing) is projected to increase by ≥15 percentage points, supported by structured prompt journaling that models a transition away from direct answer solicitation. Fifth, the Overreliance Index (proportion of AI interactions occurring before a documented individual attempt) is expected to decrease from an estimated baseline near 60% to approximately 35–40% through systematic use of Pre-Reasoning Sheets and peer review checkpoints. Sixth, log data are projected to show a rise in mean Prompt Specificity rubric scores from about 2.1 to 3.0 following iterative peer coaching and exemplar curation. Seventh, the share of deep reflective artifacts (Levels 3–4: causal explanation plus articulated transfer strategy) is expected to expand substantially as structured reflection templates routinize epistemic self-monitoring. Eighth, early teacher interviews will likely emphasize anxiety over audit workload, but mid-phase narratives are projected to pivot toward an orchestration identity aligned with reported pedagogical role evolution (Alshehri, 2023; Lee *et al.*, 2023; Karataş *et al.*, 2024). Ninth, meaningful disparities in AI literacy outcomes between urban and semi-rural cohorts are not anticipated (effect size Δ < 0.15 SD) provided curated offline pre-generated response bundles sustain parity of content exposure, thereby mitigating secondary digital divide risks. Tenth, ethical and privacy guardrails—automatic screening, pseudonymization, bias discussion prompts, and attribution scaffolds—are projected to constrain privacy-related intervention incidents to ≤2% of logged interactions. The projections described above will later be replaced with empirical estimates following inferential statistical testing (*e.g.*, ANCOVA, mediation/path modeling), robustness checks, and qualitative triangulation during the final reporting phase.

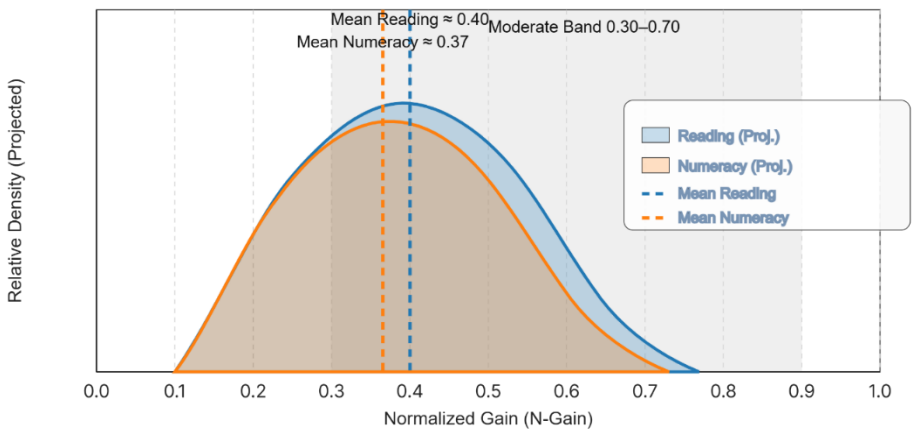

Figure 5. Projected Distribution of Reading vs Numeracy N-Gain

Figure 5 visualizes the projected (simulated) distributions of normalized gain (N-Gain) for reading and numeracy, each expected to cluster in the middle performance range (0.30–0.70) with reading centered slightly higher (≈0.40) than numeracy (≈0.37). The middle range shading emphasizes the target zone of improvement consistent with previous studies reporting decent improvement with structured verification and rapid scaffolding. The reading curve appears slightly steeper, reflecting the expectation of somewhat reduced variance due to the tighter relationship between vocabulary differentiation and explicit fact-checking cycles. The wider left shoulder for numeracy reflects the anticipated heterogeneous underlying competencies in multi-step contextual problems. The vertical dashed lines represent projected means; the final empirical curves may shift depending on the appropriateness of pre-reasoning implementation, infrastructure parity, and unanticipated ceiling or floor effects. All values remain preliminary and will be recalibrated after formal inferential analysis and assumption checks (*e.g.*, normality, homogeneity) on the collected data.

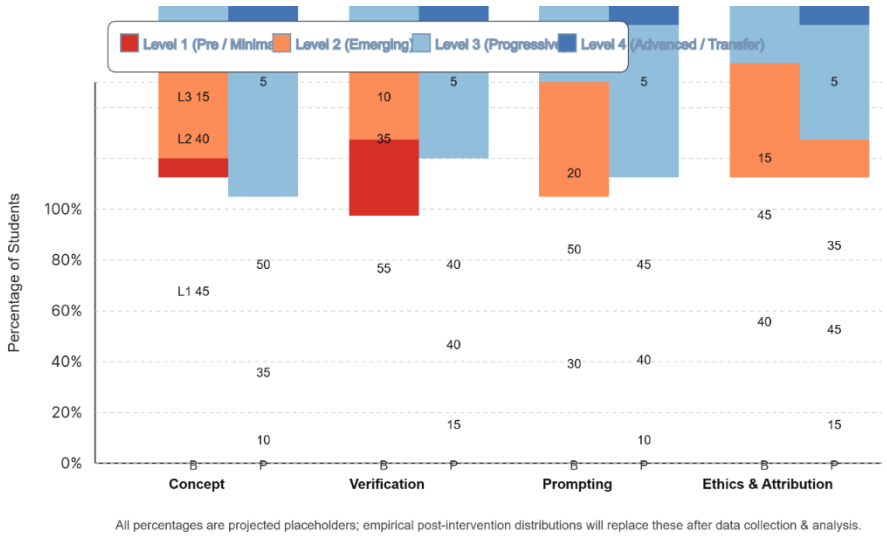

Figure 6. Anticipated Shift in AI Literacy Levels

The figure illustrates the projected redistribution of students' AI literacy levels (Levels 1–4) across four assessed dimensions: Concepts (understanding of model reasoning), Verification (ability to corroborate AI output), Prompting (precision and intentionality of prompt construction), and Ethics & Attribution (awareness of bias, privacy, and citation norms). Across all dimensions, the intervention is expected to suppress the proportion of learners at Levels 1–2 and expand Level 3 ("Progressive"), with the earliest emergence (≈5%) of Level 4 behaviors (transferable, self-initiated strategy adaptation). Verification shows the slowest shift (higher residual Level 1–2 aggregate) reflecting the greater cognitive and procedural load of multi-source reinforcement. Prompting shows a relatively larger increase to Level 3, consistent with the provision of examples and peer coaching cycles decreasing procedural complexity. Ethics & Attribution maintains a larger Level 2 segment, indicating the anticipated persistence of partially procedural (rule-following) understanding before maturing into principled transfer (Level 4). The distributions remain preliminary; actual post-test patterns will depend on implementation fidelity (especially the application of pre-AI reasoning and verification logs), infrastructure equivalence in semi-rural settings, and potential attenuation of novelty over time. Subsequent inferential analyses (*e.g.*, McNemar's or marginal homogeneity tests for level shifts, ordinal logistic modeling for predictors) will validate or revise these projections. Monochrome or journal-style variants, or versions that break down confidence bands, are available if needed.

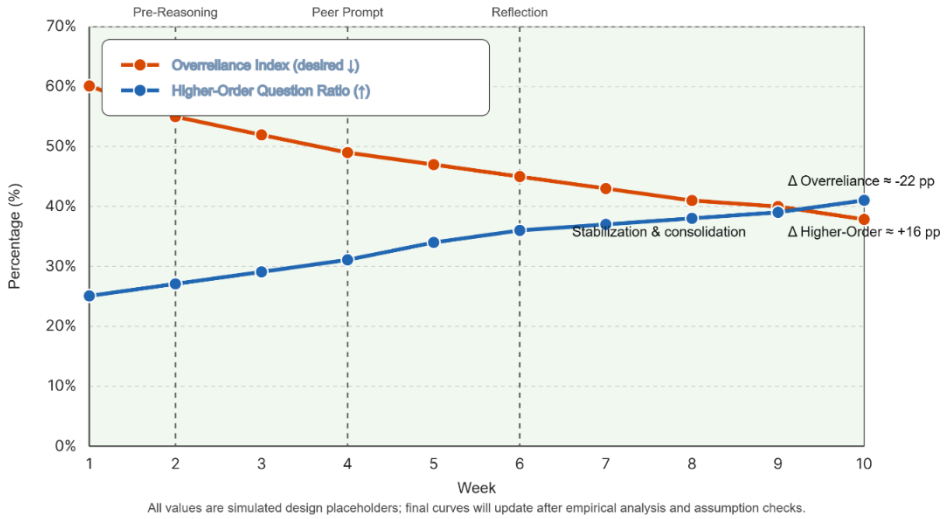

Figure 7. Decline of Overreliance Index and Rise in Higher-Order Question Rasio

The figure displays the projected weekly trajectories of two key indicators: (1) the Overreliance Index (proportion of AI interactions before documented independent attempts) which is projected to decrease from approximately 60% (Week 1) to ~38% (Week 10); (2) the High-Order Question Ratio (analysis, generalization, evaluative, in-depth clarification) which increases from ~25% to ~41% over the same period. The sharpest decrease in Overreliance occurs in the initial phase (M1–M4) in conjunction with the sequential implementation of interventions: Pre-Reasoning Sheets (M2) requires students to write down initial thoughts; Peer Prompt Coaching (M4) reduces the need to "guess" prompts; Structured Reflection Template (M6) stabilizes verification and metacognition habits so that the curve flattens (healthy plateau). The increase in the high-order question ratio indicates a shift in orientation from asking for direct answers to conceptual and evaluative exploration; the rate of acceleration increases after M5 as students begin to internalize specific prompt patterns and source verification processes. Final projected deltas: -22 percentage points for Overreliance and +16 percentage points for High-Level Questions—both within the realistic range of moderate change without relying on full cognitive substitution. All values serve as design placeholders: actual results may shift due to variations in teacher fidelity (enforcement of pre-reasoning sheets), device availability (verification cycles may be truncated in semi-rural areas), or fading novelty effects. Final analysis will include simple time-series modeling (*e.g.*, segmented regression to detect breakpoints at M2, M4, M6) and exploratory mediation tests (Prompt Quality → Verification Rigor → Overreliance ↓ → High-Level Question Ratio ↑).

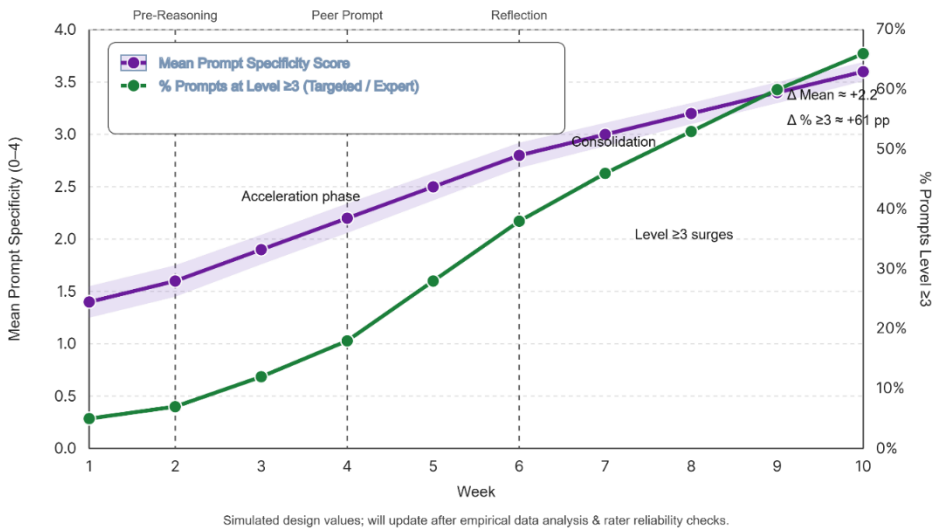

Figure 8. Trajectory of Prompt Specificity Scores Across Weeks

The figure shows projected improvement in prompt specificity quality across 10 weeks using two complementary indicators: (1) Mean Prompt Specificity Score (0–4 rubric) moving from ~1.4 (general, low-

context prompts) toward ~3.6 (context-rich prompts with format and evaluative criteria); (2) The share of prompts rated Level ≥3 (Targeted or Expert) rising sharply from 5% to 66%. The purple band depicts a design-stage variation envelope (not an empirical confidence interval), narrowing (±0.15 → ±0.09) to represent expected stabilization as structures become routinized.

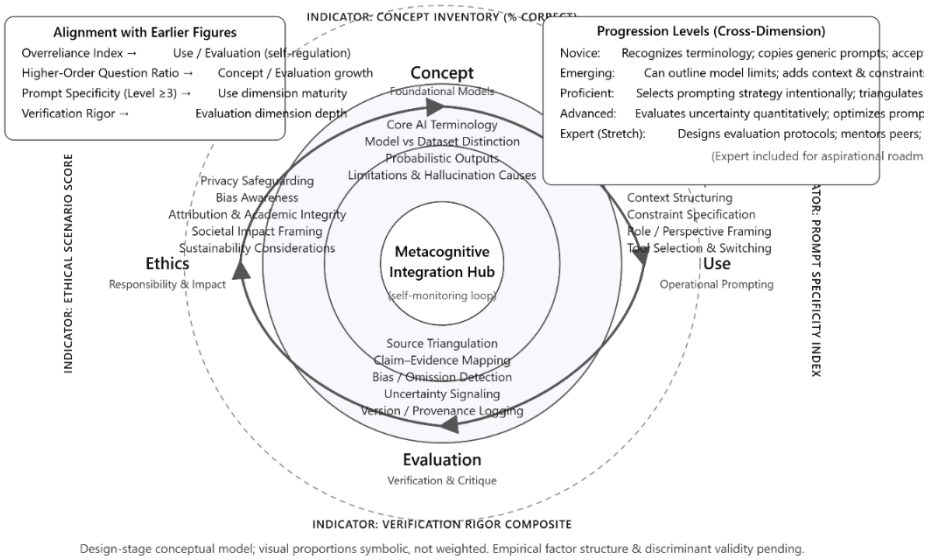

Figure 9. Framework of AI Literacy Dimensions: Concept–Use–Evaluation–Ethics

The framework articulates AI Literacy as an integrated, iterative ecosystem spanning four interdependent dimensions encircling a metacognitive integration hub. Rather than treating "AI skills" as a flat checklist, it models a dynamic learning loop: conceptual understanding informs purposeful use; applied use generates artifacts that require evaluation; evaluation yields feedback that reshapes conceptual mental models—all while ethical reasoning serves as a continuous guard-rail shaping acceptable boundaries, escalations, and governance.

3.2 Discussion

The proposed hybrid model conceptualizes AI as a "probabilistic scaffold" rather than an authoritative answer engine, aligning with learner-centered and epistemic vigilance perspectives that emphasize elaboration over cognitive outsourcing (Alshehri, 2023; Lee *et al.*, 2023). The enforced sequence—Prompt → Verification → Revision → Reflection—intentionally inserts productive friction, compelling students to interrogate AI outputs against textbooks, manual calculations, or simple empirical checks, thereby strengthening internal error schemas and skepticism. The projected moderate gains in literacy and numeracy are theoretically coherent because the design preserves substantial analog cognitive labor (manual reasoning, peer dialogue), reducing the risk of social interaction displacement. Anticipated mediating mechanisms include: increased prompt specificity narrowing output variance and lowering downstream verification burden; systematic manual verification cultivating patterned detection of anomalies; and causal reflection stabilizing transferable strategies for non-AI tasks such as independent essay composition or open-form quantitative reasoning. The reduction in overreliance reclaims metacognitive bandwidth, fostering an upward cycle in higher-order questioning depth. Teacher role evolution from routine exercise dispenser to metacognitive diagnostician and orchestration specialist converges with emerging accounts of pedagogical identity reconfiguration in AI-mediated classrooms (Karataş *et al.*, 2024; Lee *et al.*, 2023). Implementation risks persist: variability in fidelity of enforcing pre-AI attempts; infrastructural constraints in semi-rural contexts that may truncate full Prompt–Verification–Reflection cycles; and the absence of nationally codified AI literacy standards, necessitating provisional rubrics vulnerable to interpretive drift despite calibration efforts. Local contextualization (*e.g.*, traditional market pricing, water resource issues, local flora) functions as a semantic anchor rather than decorative motif, mitigating superficial exchanges and supporting durable conceptual retention (Anggreni *et al.*, 2023). Deliberate integration of character education (honest attribution, data prudence) and bias dialogue aims to shift student orientation from instrumental efficiency toward socially responsible stewardship of generative systems (Damanhuri & Wali, 2024; Rainey *et al.*, 2021). Key limitations warrant vigilance: novelty effects may inflate early engagement; asynchronous device availability could attenuate iterative cycles; and performance assumptions about a GPT-5

class model remain hypothetical, with potential variance in hallucination profiles or domain bias. Consequently, end-phase analyses should incorporate temporal segmentation (early vs late sessions) for novelty decay detection and moderator analyses based on resource tier. If projected patterns are confirmed—especially mediation paths from Prompt Quality through Verification Rigor to learning gains—findings could inform drafting of a foundational national AI literacy standard (Concept–Use–Evaluation–Ethics) aligned with holistic learner profiles, while structuring a tiered micro-credential pathway for teachers (awareness → applied auditing → orchestration). Policy implications would prioritize differential infrastructure support (adaptive offline bundles), child-specific AI privacy guidelines exceeding generic consent frameworks, and open-access themed module repositories to lower marginal adoption costs. Future research priorities include longitudinal retention of verification competencies, multimodal (image/audio) extensions for early literacy scaffolding, systematic audits of cultural narrative bias, co-design with indigenous communities to prevent semantic homogenization, and structural mediation modeling (Prompt Quality → Verification Rigor → Overreliance Reduction → Higher-Order Inquiry → Achievement Gains). Overall, the design balances generative adaptivity with pedagogical integrity via structured pre-reasoning, explicit verification, reflective abstraction, and ethical guardrails; ultimate scalability will depend on sustained teacher fidelity, robust infrastructural equity, and empirically validated causal mechanisms.

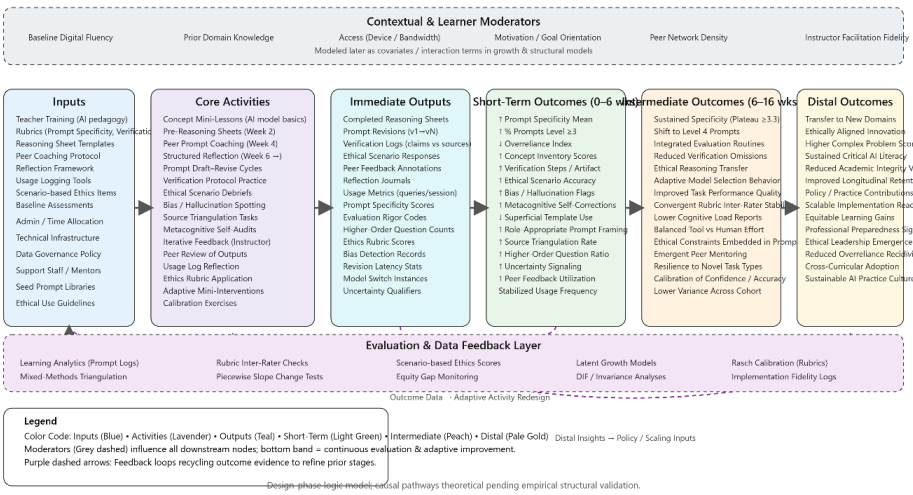

Figure 10. Logic Model Inputs to Distal Outcomes

This logic model delineates the hypothesized causal pathway from foundational resources (Inputs) and enacted pedagogical processes (Activities) through directly observable artifacts (Immediate Outputs) to multi-stage learning transformations (Short-Term → Intermediate → Distal Outcomes). It embeds moderation and feedback layers to emphasize adaptive, data-informed refinement rather than a fixed linear pipeline.

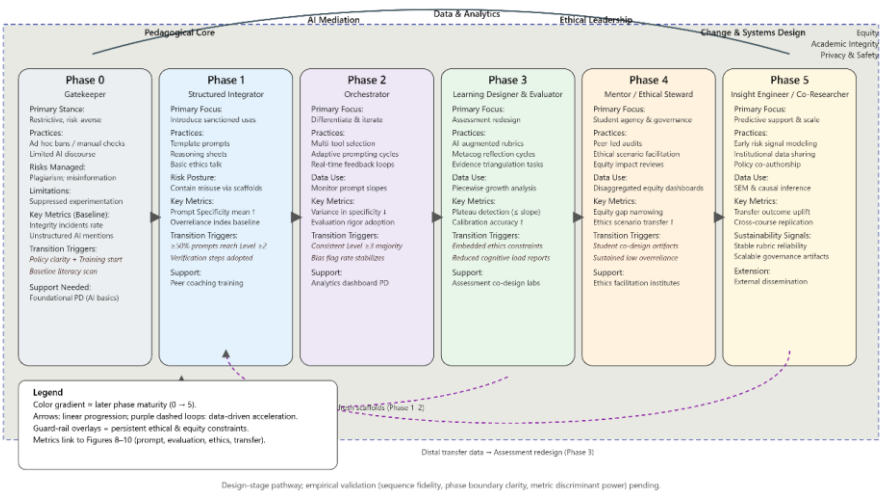

Figure 11. Teacher Role Evolution Pathway

This pathway frames teacher role evolution in AI-integrated learning environments as a staged expansion of responsibilities, data sophistication, and ethical stewardship—shifting from defensive gatekeeping to proactive, insight-generating leadership. Rather than a rigid ladder, it is a developmental continuum with feedback loops that can accelerate or stall transitions.

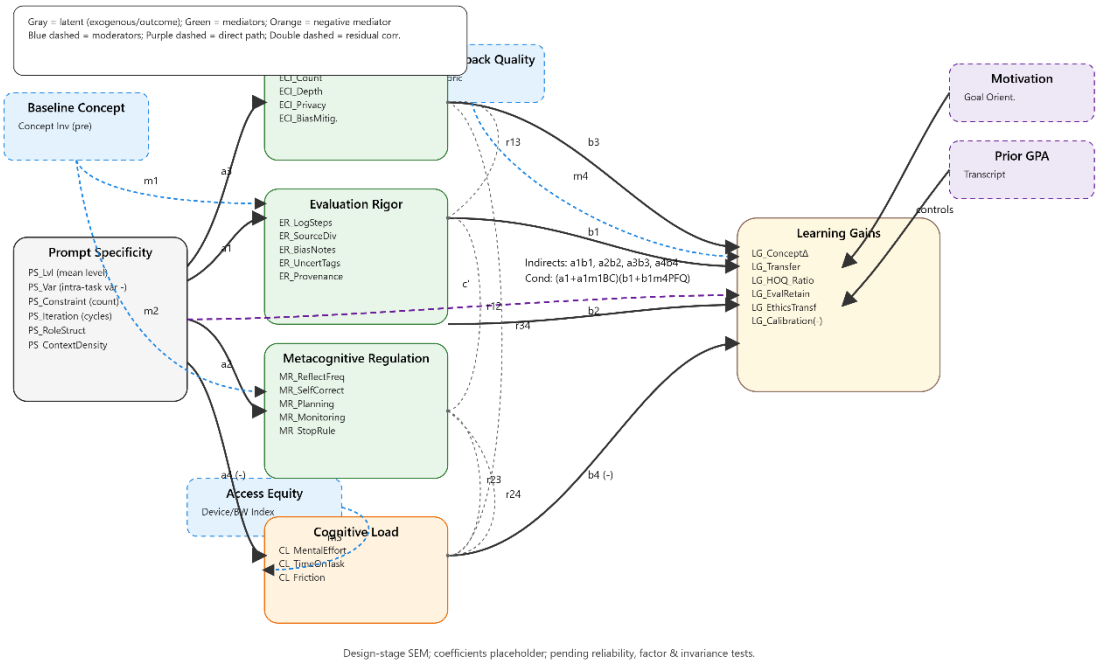

Figure 12. Proposed Mediation Model Linking Prompt Specificity to Learning Gains

This proposed structural mediation model hypothesizes that advancements in Prompt Specificity (PS) catalyze Learning Gains (LG) largely through three constructive parallel channels—Evaluation Rigor (ER), Metacognitive Regulation (MR), and Ethical Constraint Integration (ECI)—while simultaneously reducing or bypassing detrimental Cognitive Load (CL) that can suppress gains if orchestration is inefficient. Moderation terms posit conditional process dynamics: baseline conceptual knowledge strengthens the translation of specificity into evaluation and self-regulation; high-quality peer feedback amplifies the effect of evaluation rigor on distal learning; equitable access attenuates load burdens for learners with initial infrastructure disadvantages. Covariates (Motivation, Prior GPA) partial out global variance to sharpen estimates of mechanistic pathways. The proposed hybrid "probabilistic scaffold" model positions AI not as an answer machine, but as an elaboration catalyst that forces a Prompt → Verification → Revision → Reflection cycle with productive friction to maintain analog cognitive work (manual reasoning, peer dialogue) and foster epistemic understanding. The projected mediating pathway—Prompt Specificity enhancing Evaluation Rigor, Metacognitive Regulation, and Ethical Constraint Integration while simultaneously suppressing burden through Cognitive Load management—forms a testable causal framework to distinguish substantial performance improvements from illusions of automation. Moderators (foundational concepts, peer feedback quality, equity of access) open up differential orchestration, while ethical integration (honest attribution, bias, privacy) shifts students' orientation from instrumental efficiency to responsible stewardship. The key theoretical contribution lies in the paradigm shift of AI from "answer provider" to "probabilistic structure strengthening metacognitive regulation and evidence evaluation," integrating ethical dimensions with cognitive performance within a single mediating framework. In practice, granular indicators (PS, ER, MR, ECI, CL) are ready to be operationalized in a dashboard to detect stagnation or overreliance early and form the basis for designing AI literacy standards (Concept–Use–Evaluation–Ethics) and teacher micro-credential pathways (awareness → auditing → orchestration). Policy implications include differential infrastructure support (adaptive offline bundles), child privacy guidelines that go beyond general permissions, a local contextual module repository to prevent semantic homogenization, and a simple artifact verification requirement as an accountability process. Risks to anticipate include the false novelty effect on initial engagement, variability in teacher fidelity in forcing "pre-AI efforts," infrastructure limitations that can shorten the full cycle, potential interpretive drift of ethics rubrics without periodic calibration, and the resulting performance of GPT-5 classroom models (hallucinatory profiles or domain bias). Therefore, the final analysis

requires temporal segmentation (early vs. late) to map the decay of novelty effects, moderators based on resource levels, and advanced chain mediation modeling (Prompt Quality → Verification Rigor → Overreliance Reduction → Higher-Order Inquiry → Achievement). Additional research directions include longitudinal retention of competency verification, multimodal expansion for early literacy, cross-cultural narrative bias audits, and co-design with indigenous communities to prevent homogenization of meaning. This design balances generative adaptability with pedagogical integrity through structured cognitive friction, explicit verification, reflective abstraction, and integrated ethical guardrails. Its scalability is determined by three pillars: the fidelity of teacher orchestration in maintaining the process sequence, the equity of infrastructure that ensures each student can complete the full cycle, and the empirical verification of the hypothesized mediation pathways. If empirical findings confirm key mediation and moderation patterns, this model provides the conditions and operational foundation for non-reductionist standardization of AI literacy—fostering a more critical, ethical, adaptive, and inclusive learning ecosystem.

4. Conclusion

At this design stage, the hybrid "probabilistic scaffold" model demonstrates moderate but strategic potential for improving reading literacy (projected N-Gain ≈0.35–0.45) and numeracy (≈0.30–0.40) through a combination of AI-assisted vocabulary differentiation consistently paired with explicit verification and multi-step contextual problem solving that still requires manual reasoning to prevent cognitive substitution. The shift in AI literacy from Levels 1–2 to predominantly Level 3 (with embryonic Level 4 in a small proportion of students) indicates early internalization of the model's probabilistic concepts and independence of source verification. The projected decrease in the Overreliance Index (≈60% → 35–40%), the increase in the share of high-order questions (+≥15 p.p.), the jump in the average Prompt Specificity score (≈2.1 → 3.0) and the expansion of deep reflection artifacts confirm that the Prompt → Verify → Revise → Reflect cycle has the potential to reclaim metacognitive space and shift the orientation from the demand for quick answers to conceptual and evaluative exploration. Teacher narratives are predicted to shift from audit burden anxiety toward orchestration identity, while the urban–semi-rural gap is expected to remain minimal (Δ <0.15 SD) if curated offline bundles maintain exposure parity; privacy/ethics-related incidents are projected to be low (≤2%) thanks to automated screening and attribution scaffolding. Theoretically, this design sets up a testing path for the mediation mechanism (Prompt Specificity → Evaluation Rigor / Metacognitive Regulation / Ethical Constraint Integration → Learning Gains with Cognitive Load management as a negative buffer) to distinguish authentic improvements from pseudo-automation effects. However, all figures are hypothetical and susceptible to teacher fidelity variability, novelty effects, and infrastructure limitations that could truncate the full cycle; furthermore, the performance assumptions of the GPT-5 classroom model remain unvalidated (hallucination profiles or domain biases may shift). Thus, ultimate success will be determined by (1) consistent application of pre-reasoning before AI interactions, (2) equity of access that allows for complete verification and reflection for all students, and (3) empirical verification of the mediation pathway and moderators (prior knowledge, peer feedback quality, resource tier). If empirical data later confirms the projected pattern, these findings could form the foundation for developing integrated national AI literacy standards (Concept–Use–Evaluation–Ethics), tiered teacher micro-credential pathways, and differentiated policies for infrastructure and child privacy. They also confirm that the use of AI designed with productive friction and ethical safeguards can improve learning quality without encouraging passive dependency.

References

Ali, M. A., & Mughari, S. (2025). Effect of AI literacy on online information search competencies among medical students in Pakistan. *Information Development*, *41*(3), 811-824. *Information Development*. https://doi.org/10.1177/02666669241299765

Al-Qerem, W., Eberhardt, J., Jarab, A., Al Bawab, A. Q., Hammad, A., Alasmari, F., ... & Al-Beool, S. (2023). Exploring knowledge, attitudes, and practices towards artificial intelligence among health professions' students in Jordan. *BMC Medical Informatics and Decision Making*, *23*(1), 288. https://doi.org/10.1186/s12911-023-02403-0

Alshehri, B. (2023). Pedagogical paradigms in the AI era: Insights from Saudi educators on the long-term implications of AI integration in classroom teaching. *IJESA, 2*(8), 159-180. https://doi.org/10.59992/ijesa.2023.v2n8p7

Anggreni, P., Wulandari, T., Winata, E. Y., Irmawati, I., Wali, M., Arifin, N. Y., ... & Asrial, A. (2023). *PENDIDIKAN UNTUK SEMUA: Menemukan Solusi Inklusif dalam Lingkungan Pembelajaran*. Yayasan Literasi Sains Indonesia.

Biagini, G. (2024). Assessing the assessments: Toward a multidimensional approach to AI literacy. *Media Education, 15*(1), 91-101. https://doi.org/10.36253/me-15831

Brummelen, J., Heng, T., & Tabunshchyk, V. (2021). Teaching tech to talk: K-12 conversational artificial intelligence literacy curriculum and development tools. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(17), 15655-15663. https://doi.org/10.1609/aaai.v35i17.17844

Chan, K., & Zary, N. (2019). Applications and challenges of implementing artificial intelligence in medical education: Integrative review. *JMIR Medical Education, 5*(1), e13930. https://doi.org/10.2196/13930

Damanhuri, & Wali, M. (2024). How does the influence of moral values affect children's use of Roblox in character education? *International Journal Education and Computer Studies (IJECS), 4*(3), 130-142. https://doi.org/10.35870/ijecs.v4i3.3627

Druga, S., Vu, S., Likhith, E., & Qiu, T. (2019). Inclusive AI literacy for kids around the world. *Proceedings of FabLearn 2019*, 104-111. https://doi.org/10.1145/3311890.3311904

Karataş, F., Eriçok, B., & Tanrikulu, L. (2024). Reshaping curriculum adaptation in the age of artificial intelligence: Mapping teachers' AI-driven curriculum adaptation patterns. *British Educational Research Journal, 51*(1), 154-180. https://doi.org/10.1002/berj.4068

Kim, S., Kim, S., Kim, H., & Lee, Y. (2025). Integrating artificial intelligence into medical curricula: Perspectives of faculty and students in South Korea. *Korean Journal of Medical Education, 37*(1), 65-70. https://doi.org/10.3946/kjme.2025.324

Lee, J., Wu, A., Li, D., & Kulasegaram, K. (2021). Artificial intelligence in undergraduate medical education: A scoping review. *Academic Medicine, 96*(11S), S62-S70. https://doi.org/10.1097/acm.0000000000004291

Lee, Y., Kim, S., Lee, Y., Kim, H., Seo, S., Kim, H., … & Kim, K. (2023). Defining medical AI competencies for medical school graduates: Outcomes of a Delphi survey and medical student/educator questionnaire of South Korean medical schools. *Academic Medicine, 99*(5), 524-533. https://doi.org/10.1097/acm.0000000000005618

Moldt, J., Festl-Wietek, T., Mamlouk, A., Nieselt, K., Fuhl, W., & Herrmann-Werner, A. (2023). Chatbots for future docs: Exploring medical students' attitudes and knowledge towards artificial intelligence and medical chatbots. *Medical Education Online, 28*(1). https://doi.org/10.1080/10872981.2023.2182659

Nurbaya, Azwar, I., Emiliasari, R. N., Susilawati, Dewi, S. L., Rahmasari, S. M., Wali, M., Nasution, A., Inayah, S., Angraini, L. M., Kania, N., Pamungkas, M. D., Supriyadi, E., & Widyastika, D. (2023). *Inovasi Pembelajaran*. CV. Edupedia Publisher.

Park, W., & Kwon, H. (2023). Implementing artificial intelligence education for middle school technology education in Republic of Korea. *International Journal of Technology and Design Education, 34*(1), 109-135. https://doi.org/10.1007/s10798-023-09812-2

Rainey, C., O'Regan, T., Matthew, J., Skelton, E., Woznitza, N., Chu, K., … & Malamateniou, C. (2021). Beauty is in the AI of the beholder: Are we ready for the clinical integration of artificial intelligence in radiography? An exploratory analysis of perceived AI knowledge, skills, confidence, and education perspectives of UK radiographers. *Frontiers in Digital Health, 3*. https://doi.org/10.3389/fdgth.2021.739327

Rathore, A., Sultana, N., Zareen, S., & Ahmed, A. (2023). Artificial intelligence and curriculum prospects for elementary school. *Pakistan Journal of Humanities and Social Sciences, 11*(4). https://doi.org/10.52131/pjhss.2023.v11i4.1909

Sukmawati, W., & Wahjusaputri, S. (2024). Integrating RADEC model and AI to enhance science literacy: Student perspectives. *Jurnal Penelitian Pendidikan IPA, 10*(6), 3080-3089. https://doi.org/10.29303/jppipa.v10i6.7557

Weidener, L., & Fischer, M. (2023). Artificial intelligence teaching as part of medical education: Qualitative analysis of expert interviews. *JMIR Medical Education, 9*, e46428. https://doi.org/10.2196/46428.

How Cites

Yusmaneli, Y., Fathurrahmad, F., & Raihani, A. (2025). A Hybrid "Probabilistic Scaffold" Model for Enhancing Reading Literacy, Numeracy, and AI Literacy among Indonesian Primary School Students. *Design Journal*, *3*(2), 78–98. https://doi.org/10.58477/dj.v3i2.320.

Publisher's Note

Yayasan Pendidikan Mitra Mandiri Aceh (YPPMA) remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Submit your manuscript to YPMMA Journal and benefit from: https://journal.ypmma.org/index.php/dj.