JOLLS: Journal of Language and Literature Studies
Website: https://journal-center.litpam.com/index.php/jolls/index
Email: jolls@litpam.com
DOI: https://doi.org/10.36312/jolls.v6i3.5041

September 2026. Vol. 6, No. 3
e-ISSN: 2808-1099
pp. 1087-1099

The GenAI-Prompting Learning Model (GPLM): Developing and Testing a Structured Prompting Framework to Support Pre-service PBSI Teachers' Metacognitive Skills

[1] *Lisdwiana Kurniati, [1]Amy Sabila, [1]Roro Dwi Astuti
[1] Indonesian Language and Literature Education, Universitas Muhammadiyah Pringsewu (UMPRI)
Jln. KH. Ahmad Dahlan No.112, Pringsewu Utara, Pringsewu, Lampung, Indonesia

*Corresponding Author e-mail: lisdwianakurniati@umpri.ac.id

Received: May 2026; Revised: June 2026; Accepted: August 2026; Published: September 2026

Abstract

Pre-service Indonesian Language Education (Pendidikan Bahasa dan Sastra Indonesia/PBSI) teachers often struggle to translate instructional-design theory into contextually grounded, creative lesson designs. Building on evidence of a broader theory-practice gap in Indonesian pre-service language teacher education, this study examined whether this gap is compounded by underdeveloped metacognitive regulation (planning, monitoring, evaluating) among fourth-semester PBSI students enrolled in the Indonesian Language and Literature Learning Studies course at Universitas Muhammadiyah Pringsewu (UMPRI). The study aimed to develop, validate, and empirically test the GenAI-Prompting Learning Model (GPLM), a four-level structured prompting framework (Exploratory, Analytical, Synthetic, and Evaluative Prompting) designed to support students' metacognitive regulation during instructional-design tasks. An embedded mixed-methods Research and Development (R&D) design was used, operationalized through the five ADDIE stages (Analyze, Design, Develop, Implement, Evaluate); qualitative data were nested within a pre-experimental one-group pretest-posttest quantitative design and used to explain the numerical findings. The GPLM was reviewed by three expert validators (two language-pedagogy specialists and one educational-technology specialist) and implemented with 41 fourth-semester PBSI students (28 female, 13 male) over a 14-session/14-week intervention. Instruments comprised the Metacognitive Skills Test (MST), the Instructional Design Quality Rubric (IDQR), a 15-item GPLM Practicality Questionnaire, observation sheets, digital prompting journals, and semi-structured interviews with 15 purposively selected students. Content validity, calculated with Lawshe's (1975) Content Validity Ratio across three validators, reached 0.99; because CVR critical values are conventionally tabulated for panels of five or more raters, this result is reported with the underlying calculation and interpreted as indicative rather than definitive evidence of content validity. Thematic analysis of interviews and journals identified three explanatory themes-externalization of internal thought, iterative refinement through dialogic interaction, and emerging epistemic agency. The findings showed that the GPLM is a well-received, promising instructional tool associated with improved metacognitive and instructional-design scores, rather than as proof of a causal "acceleration" effect.

Keywords: GenAI-Prompting; Instructional design; Metacognitive skills; Pre-service language teacher education

How to Cite: Kurniati, L., Sabila, A., & Dwi, R. (2026). The GenAI-Prompting Learning Model (GPLM): Developing and Testing a Structured Prompting Framework to Support Pre-service PBSI Teachers' Metacognitive Skills. *Journal of Language and Literature Studies, 6*(3), 1087-1099. doi: https://doi.org/10.36312/jolls.v6i3.5041

https://doi.org/10.36312/jolls.v6i3.5041

INTRODUCTION

The Indonesian Language Education (Pendidikan Bahasa Indonesia-PBSI) curriculum at higher-education institutions increasingly requires students not only to acquire theoretical content but also to develop Higher-Order Thinking Skills (HOTS), which fundamentally encompass metacognitive competencies (Lutfi, 2023). Metacognition-the ability to consciously plan, monitor, and evaluate one's own cognitive

processes-is particularly relevant under the Merdeka Belajar (Freedom to Learn) Curriculum, which emphasizes student agency, contextual learning design, and adaptive pedagogical innovation. The Indonesian Language and Literature Learning Studies course (Kajian Pembelajaran Bahasa dan Sastra Indonesia) at UMPRI exemplifies this challenge: fourth-semester PBSI students are expected to synthesize theories of learning models with the demands of language and literature content, yet a documented theory-practice gap persists in Indonesian pre-service language teacher education more broadly (Setiawan & Suwatno, 2024). This study addresses a narrower version of that gap: the specific difficulty PBSI students at UMPRI experience when translating instructional-design theory into classroom-ready lesson designs within a single, compulsory instructional-design-related course.

A needs analysis was conducted with the same 41 fourth-semester PBSI students who later participated in the intervention, using structured classroom observation across 14 preceding sessions and the Metacognitive Skills Test (MST) that was subsequently re-administered as the study's pretest. This design choice - using the needs-analysis administration of the MST as the pretest - means the "baseline" and "pretest" figures reported in this article refer to the same measurement occasion and the same instrument; the values are therefore reported once and referenced consistently throughout the article rather than as two independent data points. The needs analysis showed that 73% of students experienced significant difficulty in the critical synthesis phase, specifically when required to transform conceptual knowledge of instructional approaches such as Discovery Learning, Problem-Based Learning, and Project-Based Learning into contextually relevant and creative instructional designs for language and literary materials. On the MST, students' planning scores averaged 62.4 out of 100, while monitoring and evaluating scores were lower, at 54.8 and 51.2 respectively. These findings indicate that students predominantly adopted a cookbook approach - replicating existing templates without engaging in reflective, strategic design thinking - underscoring the need for a structured cognitive-scaffolding framework (Zimmerman & Schunk, 2019).

The rapid emergence of Generative Artificial Intelligence (GenAI) has opened new possibilities for cognitive augmentation in education. Platforms such as ChatGPT and Gemini are no longer limited to information retrieval; recent work frames interaction with such systems as imposing genuine metacognitive demands on users - requiring monitoring and control over how prompts are formulated, how outputs are evaluated, and how workflows are adjusted (Tankelevitch et al., 2024). This reframes the practice this study calls GenAI-Prompting as something distinct from several related but non-identical constructs. GenAI-Prompting, as operationalized in the GPLM, refers specifically to a staged, pedagogically sequenced use of AI dialogue - moving learners through exploratory, analytical, synthetic, and evaluative prompting acts - for the explicit purpose of externalizing and training planning, monitoring, and evaluating. It differs from prompt engineering, which is typically technique-focused and oriented toward optimizing an AI system's output quality rather than the user's own cognitive regulation (Gupta & Bhardwaj, 2023); from AI literacy, which is a broader competency encompassing technical, ethical, and critical-evaluation knowledge about AI systems rather than a specific instructional sequence (Darwin et al., 2024); from dialogic prompting or self-questioning strategies used independently of AI (King, 1991), which do not involve an external, responsive interlocutor; and from ordinary, unstructured use of tools such as ChatGPT or Gemini for information retrieval, which the needs analysis found students already engaged in without measurable metacognitive benefit.

Conceptually, the GPLM connects five bodies of theory into a single instructional architecture. Metacognition (Flavell, 1979) supplies the target constructs - planning, monitoring, evaluating. Self-regulated learning theory (Zimmerman, 2000) explains how externally scaffolded regulation can be gradually internalized. Vygotskian scaffolding

(Vygotsky, 1978) explains why a structured, staged sequence of prompting levels - rather than open-ended AI use - is expected to support performance within students' zone of proximal development. Cognitive load theory (Sweller, 1988) explains why an unfamiliar prompting routine may initially increase extraneous load before practice reduces it. Feedback theory (Hattie & Timperley, 2007) explains why the near-immediate, iterative nature of AI dialogue may function differently from delayed human feedback. Read together, these five strands generate a specific hypothesis - that a staged prompting sequence, rather than GenAI access alone, is the active ingredient supporting metacognitive regulation - which this study tests empirically.

Existing scholarship on GenAI in education has examined effects on academic writing quality, student engagement, and general AI literacy (Sallam et al., 2024; Darwin et al., 2024), and a growing literature has begun to connect GenAI use to self-regulated learning and metacognitive support more directly (Tankelevitch et al., 2024). However, the specific intersection of structured, staged prompting frameworks and metacognitive skill development among pre-service language teachers engaged in instructional-design tasks remains comparatively underexplored. The claim that the GPLM is entirely without precedent is not made here; rather, this study positions the GPLM as a domain-specific contribution - one that operationalizes staged prompting explicitly around planning, monitoring, and evaluating within Indonesian language teacher education - building on, rather than superseding, the broader prompting, AI-literacy, and metacognitive-scaffolding literatures reviewed above. A systematic comparison with these adjacent frameworks is presented in the Discussion section.

On the basis of the gaps identified above, this study addresses three interrelated research questions: (1) What are the needs and defining characteristics of a GenAI-Prompting-based learning model relevant to the Indonesian Language and Literature Learning Studies course at PBSI UMPRI? (2) How can the GPLM be designed to achieve acceptable content validity and practicality for classroom use? (3) To what extent are changes in metacognitive skills and instructional-design quality observed among fourth-semester PBSI students following GPLM implementation? Consistent with the pre-experimental, one-group design, these questions are framed in terms of observed association and change rather than causal acceleration; the study does not test, and does not claim to demonstrate, that the GPLM causes metacognitive change relative to an untreated comparison group.

The purpose of this article is therefore threefold: to report the developmental process and expert validation of the GPLM; to present empirical evidence of change in metacognitive and instructional-design scores associated with a 14-session implementation among 41 PBSI students; and to offer theoretical and practical considerations for integrating Generative AI as a metacognitive scaffolding tool in Indonesian language teacher education, framed with appropriate caution given the study's design limitations.

METHOD

Research Design

This study used an embedded mixed-methods design (Creswell & Plano Clark, 2011) nested within a Research and Development (R&D) framework operationalized through the ADDIE model (Analyze, Design, Develop, Implement, Evaluate; Branch, 2009). Concretely, the quantitative strand - a pre-experimental one-group pretest-posttest design - constituted the primary framework for testing change in metacognitive and instructional-design scores, while the qualitative strand (observation, interviews, digital journals) was embedded within the Implement and Evaluate stages to explain how and why the observed quantitative changes occurred. Integration followed a connecting strategy: quantitative results were analyzed first, and qualitative themes were then used

specifically to interpret and contextualize those results, with points of integration made explicit in the Results and Discussion section. This structure follows Branch's (2009) argument that ADDIE provides a systematic, iterative process supporting both the theoretical validity and practical implementability of educational products, and Creswell and Plano Clark's (2011) rationale that combining quantitative evidence of how much change occurred with qualitative evidence of why and how it occurred yields a more complete account than either strand alone.

Research Participants

The population comprised all students enrolled in the PBSI programme at UMPRI. The sample consisted of the 41 fourth-semester PBSI students enrolled in the Indonesian Language and Literature Learning Studies course during the even semester of Academic Year 2024/2025 (28 female, 13 male; age range 19-22 years). Total sampling was used because the course section is a compulsory, intact class; this provided an ecologically valid context and avoided selection bias, but it also means, as noted in the Limitations, that the sample was not randomly drawn from the wider PBSI population and generalization beyond similar cohorts should be made cautiously (Fraenkel, Wallen, & Hyun, 2012). All participants reported prior familiarity with GenAI tools (e.g., ChatGPT, Gemini) for general information retrieval, but none had received formal instruction in structured prompting prior to the intervention. The same 41 participants completed both the pretest and posttest; there was no separate control group, and pretest scores are not treated as a control condition in this article.

Research Instruments

Table 1 lists the instruments used, replacing the mislabeled table in the original submission (which in fact reported pre- and post-intervention IDQR results and has been relocated to the Results section as Table 2).

Table 1. Research Instruments, Constructs, and Data Sources

Instrument	Type / Items	Construct Measured	Timing	Respondents / Source
Metacognitive Skills Test (MST)	Open-ended essay; 3 tasks (one per dimension)	Planning, monitoring, evaluating (Schraw & Dennison, 1994)	Pretest (Week 1) and posttest (Week 14)	All 41 students
Instructional Design Quality Rubric (IDQR)	Analytic rubric; 5 dimensions	Quality of students' lesson-plan products	Pre- and post-intervention	All 41 students' design products
GPLM Practicality Questionnaire	15-item, 5-point Likert; 3 dimensions (5 items each)	Clarity, relevance, ease of integration	Post-intervention (Week 14)	All 41 students
Observation Sheet	Structured checklist; indicators of prompting behaviour and on-task engagement per session	Student-AI interaction pattern	Weekly, Sessions 1-14	Lecturer/observer
Digital Prompting Journal	Logged prompt-response transcripts with a brief reflective note	Prompting iteration and content of dialogic exchanges	Continuous, Sessions 1-14	All 41 students
Semi-structured Interview	12-item protocol	Explanatory account of metacognitive experience	Post-intervention (Week 14-15)	15 purposively selected students

The IDQR assessed five dimensions of students' final lesson-plan outputs: alignment with learning outcomes, appropriateness of the selected instructional model, clarity of learning steps, assessment quality, and pedagogical innovation. Pre- and post-intervention products were scored by the course lecturer using the same rubric at both occasions; raters were not blinded to assessment occasion, which is acknowledged as a limitation of the design (see Limitations). The GPLM Practicality Questionnaire comprised 15 items across three five-item dimensions - clarity of prompting instructions, relevance to instructional-design tasks, and ease of integration into existing learning habits - content-validated by the same three expert validators described below, with reliability of alpha = [author to insert Cronbach's alpha]. Mean scores were interpreted using conventional five-point category bands (1.00-1.79 very low, 1.80-2.59 low, 2.60-3.39 moderate, 3.40-4.19 high, 4.20-5.00 very high). Interview participants (n = 15) were purposively selected by ranking pretest MST scores into tertiles and drawing five students from each band (high-, middle-, and lower-performing), so that the qualitative sample deliberately represented the range of baseline metacognitive ability. Observation sheets used a structured checklist of prompting-behaviour indicators (e.g., prompt specificity, follow-up questioning, revision in response to AI feedback) scored per session by the lecturer/observer.

Procedure

The study followed the five ADDIE stages. In the Analyze stage, the needs analysis described above was conducted with the same 41 participants, together with a curriculum review of the Merdeka Belajar framework and course competency standards, and a literature review on metacognitive learning theory (Schraw & Dennison, 1994; Veenman et al., 2006) and GenAI pedagogical integration (Luckin et al., 2016). In the Design stage, the GPLM's four hierarchical prompting levels were developed - (1) Exploratory Prompting, in which students query AI to generate multiple alternative instructional approaches; (2) Analytical Prompting, in which students prompt AI to compare the advantages and limitations of each approach; (3) Synthetic Prompting, in which students instruct AI to help integrate selected elements into a coherent lesson design; and (4) Evaluative Prompting, in which students prompt AI to identify inconsistencies, gaps, or weaknesses in the completed design - and the MST, IDQR, and Practicality Questionnaire were drafted.

In the Develop stage, the GPLM instructional module, student prompting guide, and lecturer handbook were produced and submitted to three expert validators (two specialists in Indonesian Language and Literature pedagogy, one specialist in educational technology and AI applications), who rated each component on a five-point scale against construct validity, content validity, and implementability criteria. Each of the [author to insert number] items in the validation instrument was rated individually by each validator as "essential," "useful but not essential," or "not necessary" following Lawshe (1975); item-level CVR was calculated as CVR = (ne - N/2) / (N/2), where ne is the number of validators rating an item "essential" and N = 3 is the panel size. With N = 3, an item is only classifiable as retained if all three validators (ne = 3) rate it essential, which yields CVR = 1.00 for retained items; the reported aggregate CVR of 0.99 is the mean of item-level CVR values across all items after two items were revised and one item removed following validator feedback (see Table [author to insert] for the full item-by-item breakdown). Because Lawshe's original critical-value table begins at a panel of five raters, this aggregate value is reported transparently rather than presented as meeting a conventionally tabulated significance threshold, and readers should treat it as descriptive evidence from a small expert panel rather than as a definitive validity index.

In the Implement stage, the validated GPLM was delivered over 14 weekly 100-minute sessions within the Indonesian Language and Literature Learning Studies course.

Students used free-tier accounts on ChatGPT (GPT-4o mini or comparable free model, depending on availability during the study period) and Google Gemini, accessed via personal devices and campus internet; no institutional or paid-tier accounts were used. Each session followed a common structure: a short lecturer-led framing of the week's instructional-design topic, individual or paired GenAI-Prompting practice using structured templates aligned to that week's target prompting level (provided in the student prompting guide), and a closing whole-class debrief in which the lecturer, acting as facilitator rather than content-deliverer, reviewed representative prompts and AI responses. Weekly assignments required students to submit their evolving lesson-plan draft together with the corresponding prompting-journal log. Treatment fidelity was monitored through the lecturer's weekly observation checklist (recording whether each session's target prompting level was practiced) and through review of digital journals; students were permitted to consult AI tools outside class sessions for the same assignments, and this outside use was also logged in their journals. When AI outputs were inaccurate or pedagogically inappropriate, students were instructed - as part of the Evaluative Prompting level - to cross-check claims against course materials and to flag unresolved discrepancies to the lecturer, who verified and corrected them during the weekly debrief; this verification procedure is a substantive part of the model's design rather than an incidental safeguard, given that GenAI outputs can be factually unreliable (Tankelevitch et al., 2024).

Data Analysis

Prior to inferential testing, the Shapiro-Wilk test was applied to the pretest and posttest distributions and, following the reviewer's recommendation, to the pretest-posttest difference scores, which is the more appropriate normality check for a paired-sample t-test. Paired pretest-posttest scores were analyzed using a paired-sample t-test (alpha = 0.05); alongside the t-statistic and p-value, the 95% confidence interval for the mean difference and Cohen's d were computed as effect-size indicators (values to be inserted by the author from the underlying dataset). Practical significance was additionally assessed using the Normalized Gain (N-Gain) formula, g = (Spost - Spre) / (Smax - Spre), interpreted as g >= 0.70 high, 0.30 <= g < 0.70 moderate, and g < 0.30 low effectiveness (Hake, 1999). Differences among the three dimension-level N-Gain values (Planning, Monitoring, Evaluating) were not themselves subjected to inferential comparison in the original analysis; where this article discusses the relative ordering of these values, it is described as a descriptive trend rather than a statistically confirmed differential effect (see Discussion). Instructional-design quality change was assessed descriptively through mean IDQR score comparison across the five dimensions. Qualitative data (interview transcripts, observation records, digital prompting journals) were analyzed using Braun and Clarke's (2006) six-phase thematic analysis: (1) data familiarization, (2) initial code generation, (3) theme searching, (4) theme review, (5) theme definition and naming, and (6) report production. Themes were developed inductively and then mapped onto the planning, monitoring, and evaluating constructs to help interpret the quantitative results.

Ethical Considerations

This study received ethical clearance from [author to insert name of ethics committee and approval number]. Written informed consent was obtained from all 41 participants prior to data collection, including consent for the collection and analysis of digital prompting-journal logs. Prompting-journal transcripts and interview data were anonymized prior to analysis, and access to raw prompt logs was restricted to the research team; any personally identifiable or otherwise sensitive content volunteered by students within their GenAI conversations was excluded from reported excerpts. [Author:

please confirm and complete this paragraph with the actual approval details, consent procedure, and data-handling protocol used in the study.

RESULTS AND DISCUSSION

Research Results

The GPLM received favourable ratings from the expert validators, students reported high practicality, metacognitive scores increased from pretest to posttest across all three dimensions, instructional-design products received higher ratings after the intervention, and thematic analysis identified three themes that help explain these quantitative changes. These findings are reported below in the order of the study's three research questions and the ADDIE stages from which they arose, followed by a separate Discussion subsection that interprets their meaning, situates them in the literature, and considers alternative explanations.

GPLM Validity: Expert Validation Findings

Three validators - two specialists in Indonesian Language and Literature pedagogy and one specialist in educational technology and AI applications - rated the GPLM's construct validity, content validity, and implementability. As detailed in the Method, the aggregate Content Validity Ratio (CVR), calculated using Lawshe's (1975) formula across the three-member panel, was 0.99 following two item revisions and one item removal; because this panel size is smaller than the minimum of five raters for which Lawshe's critical-value table is conventionally reported, this figure is interpreted as descriptive evidence of strong expert agreement within a small panel rather than as meeting a formally tabulated significance threshold. The instructional module received a mean validity score of 4.72 out of 5.00, the student prompting guide 4.68, and the lecturer handbook 4.61.

GPLM Practicality: Student Perception and Usability

Post-intervention administration of the 15-item GPLM Practicality Questionnaire to all 41 participants yielded a mean score of 4.58 out of 5.00 (SD = 0.31; alpha = [author to insert]), in the "very high" category band. Across the three five-item sub-dimensions, students rated "clarity of prompting instructions" highest (M = 4.71), followed by "relevance to instructional-design tasks" (M = 4.63) and "ease of integration into existing learning habits" (M = 4.39). Thirty-eight of 41 students (92.7%) indicated they would recommend the GPLM to peers in other PBSI cohorts.

Metacognitive Skills: Pretest-Posttest Change

Table 2 presents the pretest and posttest means for each metacognitive dimension alongside the corresponding N-Gain values.

Table 2. Pretest-Posttest Metacognitive Skills Scores and N-Gain Values (n = 41)

Metacognitive Dimension	Pretest Mean	Posttest Mean	N-Gain	Category
Planning	62.4	84.7	0.59	Moderate
Monitoring	54.8	80.3	0.56	Moderate
Evaluating	51.2	82.9	0.65	Moderate
Overall Metacognitive Score	56.1	82.6	0.60	Moderate

The Shapiro-Wilk test confirmed normality of the pretest-posttest difference scores (W = [author to insert], p = [author to insert]), in addition to the previously reported pretest (W = 0.971, p = 0.382) and posttest (W = 0.964, p = 0.227) distributions, satisfying the assumption for a paired-sample t-test. The t-test yielded $t(40) = 18.74$, $p < .001$, 95% CI [author to insert], Cohen's d = [author to insert], indicating a statistically

significant pretest-to-posttest increase in overall metacognitive scores. The overall N-Gain value of 0.60 falls within the moderate-effectiveness category (Hake, 1999). Descriptively, N-Gain was somewhat higher for Evaluating (g = 0.65) than for Planning (g = 0.59) and Monitoring (g = 0.56), and Monitoring recorded the lowest value. These dimension-level differences were not tested statistically in this study and are therefore reported here as a descriptive pattern; their possible interpretation is discussed, with appropriate hedging, in the Discussion below.

Instructional Design Quality: Pre-Post Change

Beyond metacognitive scores, IDQR ratings of students' instructional-design products also increased. Pre-intervention mean IDQR scores averaged 61.3 out of 100; post-intervention scores rose to 83.7, a 36.4% increase. Table 3 presents the mean scores across the five IDQR dimensions before and after the intervention.

Table 3. Pre- and Post-Intervention IDQR Mean Scores Across Five Dimensions (n = 41)

IDQR Dimension	Pre-Intervention	Post-Intervention	Gain (Delta)
Alignment with Learning Outcomes	63.2	85.4	+22.2
Appropriateness of Instructional Model	58.7	82.1	+23.4
Clarity of Learning Steps	64.1	84.9	+20.8
Assessment Quality	59.4	81.7	+22.3
Pedagogical Innovation	61.1	84.3	+23.2

As noted in the Method, pre- and post-intervention products were scored by the same rater (the course lecturer) using the IDQR at both occasions, without independent double-scoring or blinding to assessment occasion; this is acknowledged as a limitation on the internal validity of the reported gains.

Qualitative Findings

Thematic analysis of 15 in-depth interviews and 41 digital prompting journals yielded three themes that help explain the quantitative changes reported above: (1) Externalization of Internal Thought, (2) Iterative Refinement Through Dialogic Interaction, and (3) Emerging Epistemic Agency.

Externalization of Internal Thought was the most frequently coded theme (14 of 15 interviews). Students across ability groups described constructing a prompt as requiring them to articulate what they knew, did not know, and needed to know - a process resembling the self-questioning strategy described by King (1991). Iterative Refinement Through Dialogic Interaction was evident in the digital journals: students averaged 6.3 prompting iterations per design task across the 14 sessions, rising from an average of 3.1 in Sessions 1-3 to 8.7 in Sessions 12-14. Emerging Epistemic Agency was most pronounced among mid- and low-ability interviewees: 9 of 10 mid- and low-ability students described themselves, post-intervention, as capable of generating, testing, and revising instructional-design arguments independently, with GenAI functioning as a sounding board rather than an authority.

Discussion

The findings showed that a distinct interpretive layer, separated from the numerical results reported. Four points are addressed in turn: plausible mechanisms linking the four prompting levels to the observed changes; the study's place within recent, prompting- and metacognition-focused GenAI literature; alternative, non-causal explanations for the observed pre-post changes; and the risks and boundaries of treating GenAI as a "cognitive mirror" or "critical friend."

The four GPLM prompting levels may plausibly stimulate different metacognitive sub-processes: Exploratory Prompting may support goal setting and option generation;

Analytical Prompting may support comparison among alternatives; Synthetic Prompting may support integration of selected elements; and Evaluative Prompting may support error detection, revision, and self-explanation. These are presented as plausible pathways consistent with the qualitative themes reported above, not as proven causal mechanisms, since the study design cannot isolate the independent contribution of each prompting level from general exposure to the course or from repeated practice. Within this hedged framing, the comparatively higher descriptive N-Gain for Evaluating is consistent with Evaluative Prompting compressing the feedback loop between producing and correcting work-broadly in line with Hattie and Timperley's (2007) account of feedback timing and with recent characterizations of GenAI interaction as imposing recurring monitoring-and-control demands on users (Tankelevitch et al., 2024), but as noted above, the dimension-level differences were not statistically tested and this interpretation should be read as a plausible account of a descriptive pattern rather than a confirmed differential effect. Similarly, the comparatively lower gain for Monitoring may reflect that Analytical Prompting, structured as discrete, turn-based dialogue, is better suited to bounded comparison tasks than to the continuous, in-process regulatory attention that monitoring requires (Zimmerman & Schunk, 2019); this suggests that future GPLM iterations might usefully add real-time scaffolds such as think-aloud protocols or structured self-monitoring checklists alongside Analytical Prompting.

The findings are systematically situated within recent work on GenAI, prompting, and metacognition. Tankelevitch et al. (2024) argue conceptually that GenAI interaction imposes metacognitive demands around prompting, evaluating outputs, and adjusting workflows; the present study extends this by empirically testing whether a staged, pedagogically sequenced prompting framework - rather than unstructured GenAI use - is associated with measurable change in students' planning, monitoring, and evaluating scores within a specific instructional-design context. Darwin et al. (2024) reported EFL students' mixed perceptions of GenAI's benefits and limitations for critical thinking, using a similar higher-education language-learning population but without a structured, staged prompting intervention or a validated metacognitive instrument; this study's inclusion of the MST provides a level of construct differentiation (planning, monitoring, evaluating measured separately) that is largely absent from that literature. Kofinas (2025) examined GenAI's implications for academic integrity in authentic assessment, a concern directly relevant to the GPLM's classroom use of AI-assisted design work and taken up further below. Relative to earlier work cited in the original submission (e.g., Huang et al., 2022; Sallam et al., 2024), which examined engagement and writing-quality outcomes without isolating metacognitive dimensions or domain-specific prompting scaffolds, the present study's contribution lies in combining a staged prompting design with a construct-differentiated metacognitive measure in a specific pre-service language-teacher-education setting - a comparatively narrow but more precisely specified niche than a general claim of first-of-its-kind novelty.

This study used a one-group pretest-posttest design without a control condition, several alternative, non-causal explanations for the observed pre-post changes cannot be ruled out and should be considered alongside the prompting-based account above. These include novelty effects associated with using an unfamiliar technology in a structured way for the first time; increased lecturer attention during the 14-session intervention relative to typical instruction; general repeated-practice effects from completing multiple instructional-design tasks across 14 sessions regardless of prompting structure; assessment familiarity, since students encountered the MST and IDQR criteria at pretest before being assessed again at posttest; peer interaction during paired prompting practice; and increased motivation associated with using a novel technology. The pre-experimental design used here cannot separate these explanations from the effect of GPLM's specific prompting structure, and this is presented as a substantive limitation

rather than a minor caveat; the Recommendation section proposes a controlled design as a next step precisely to address this.

The Cognitive Mirror Hypothesis advanced in this study - that GenAI, embedded within a structured prompting framework, can function as a reflective surface that makes learners' own reasoning visible - is offered here with clearer boundaries than in the original submission. The hypothesis is intended to apply specifically to staged, pedagogically scaffolded prompting of the kind implemented in the GPLM, not to GenAI use in general; the qualitative themes reported above (externalization, iterative refinement, emerging epistemic agency) are the observable mechanisms taken as consistent with, though not conclusive proof of, this hypothesis. Treating GenAI as a "critical friend" in this way also carries risks that must be discussed rather than assumed away. AI systems can generate incorrect or unreliable feedback, and users often struggle to calibrate appropriate confidence in AI outputs (Tankelevitch et al., 2024); the model therefore risks propagating hidden bias or factual error if students accept AI evaluative feedback uncritically.

There is a related risk of overconfidence or shallow agreement, in which students adopt AI suggestions without genuine critical engagement, which would undermine rather than support metacognitive development. Repeated reliance on AI as an evaluative partner may also foster epistemic dependence if not paired with explicit training in independent judgment. Because GPLM classroom use of GenAI for design work sits close to broader academic-integrity questions about AI-assisted assessment (Kofinas, 2025), instructors implementing similar models should make explicit which forms of AI assistance are permitted and how originality is assessed. Finally, privacy concerns around prompt-log content and unequal access to reliable internet or capable devices are practical constraints on any AI-mediated pedagogical model and are addressed in this study's Ethical Considerations and Limitations.

CONCLUSION

This study developed and expert-validated the GenAI-Prompting Learning Model (GPLM), a four-level structured prompting framework, and implemented it with 41 fourth-semester PBSI students at UMPRI over a 14-session intervention within a single instructional-design-related course. In answer to the study's three research questions: the needs analysis identified a specific, course-bounded gap between instructional-design theory and practice linked to underdeveloped planning, monitoring, and evaluating skills (RQ1); the GPLM, reviewed by three expert validators, achieved a high mean validity rating and a CVR of 0.99, interpreted cautiously given the small validator panel, together with high student-rated practicality (M = 4.58/5.00) (RQ2); and metacognitive scores and instructional-design ratings were both higher at posttest than at pretest, with an overall N-Gain of 0.60 (moderate) and a statistically significant paired-sample t-test result (RQ3). Given the pre-experimental, one-group design, these results are best read as evidence that the GPLM was judged favourably by the participating experts and students and that participants' scores were higher after implementation - not as confirmation that the GPLM causally accelerated metacognitive development relative to an untreated comparison condition.

Qualitative findings from interviews and digital prompting journals suggested three plausible explanatory mechanisms - externalization of internal thought, iterative refinement through dialogic interaction, and emerging epistemic agency - that are consistent with, though not sufficient to confirm, the Cognitive Mirror Hypothesis advanced in this study. The principal limitations of this study are the absence of a control group, single-rater (non-blinded) IDQR scoring, a small expert-validator panel relative to conventional CVR reporting practice, and a single-institution, single-course sample; these limitations bound the strength of the causal and generalization claims that

can be responsibly drawn from the findings. With these caveats, the GPLM represents a promising, classroom-tested starting point for integrating staged GenAI-Prompting as a metacognitive scaffold in language teacher education, warranting further testing under more rigorous designs before broader adoption is recommended.

RECOMMENDATION

Based on the findings and limitations of this study, several recommendations are proposed. First, subsequent studies should employ a true experimental or quasi-experimental design with a control or comparison group - for example, comparing structured GPLM prompting against unstructured GenAI use - to isolate the effect of the prompting structure itself from novelty, practice, and attention effects discussed above. Second, IDQR scoring in future studies should use at least two independent raters, blinded to assessment occasion, with inter-rater reliability reported. Third, the relatively lower descriptive N-Gain observed for Monitoring suggests that future GPLM iterations should incorporate supplementary real-time monitoring tools, such as think-aloud protocols, structured self-monitoring checklists, or AI-integrated reflective dashboards, alongside Analytical Prompting. Fourth, content-validity procedures in future studies should involve a validator panel of at least five members to allow CVR results to be benchmarked against Lawshe's (1975) published critical values, or should use an alternative index such as Aiken's V, which is more commonly used with smaller expert panels. Finally, as GPLM-type models are extended to other teacher-education fields, researchers should give explicit attention to academic-integrity safeguards, AI output verification procedures, equitable access to AI tools, and data-privacy protections for prompt logs, building on the risks discussed.

ACKNOWLEDGMENT

The authors would like to express sincere gratitude to Universitas Muhammadiyah Pringsewu (UMPRI) for institutional support throughout the conduct of this research. Appreciation is also extended to the expert validators who contributed their time and expertise to the validation process, and to the forty-one fourth-semester PBSI students whose participation and engagement made this study possible.

Declaration of GenAI Use

GenAI platforms (ChatGPT and Gemini) were used as the pedagogical intervention studied in this article - that is, as the instructional medium experienced by student participants - and this use is the subject of the research reported here, not a tool used to produce the manuscript itself. All AI-generated material referenced or produced in connection with this study, whether as part of the classroom intervention or in manuscript preparation, was reviewed and verified by the human authors, who retain full responsibility for the accuracy, interpretation, originality, and citation integrity of this article.

REFERENCES

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's educational objectives. Longman.

Bandura, A. (1997). Self-efficacy: The exercise of control. W. H. Freeman.

Bereiter, C., & Scardamalia, M. (1987). The psychology of written composition. Lawrence Erlbaum Associates.

Branch, R. M. (2009). Instructional design: The ADDIE approach. Springer.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Brown, A. L. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), Metacognition, motivation, and understanding (pp. 65-116). Lawrence Erlbaum Associates.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Rand McNally.

Creswell, J. W., & Plano Clark, V. L. (2011). Designing and conducting mixed methods research (2nd ed.). SAGE Publications.

Darwin, Rusdin, D., Mukminatien, N., Suryati, N., Laksmi, E. D., & Marzuki. (2024). Critical thinking in the AI era: An exploration of EFL students' perceptions, benefits, and limitations. Cogent Education, 11(1), 2290342. https://doi.org/10.1080/2331186X.2023.2290342

Emig, J. (1977). Writing as a mode of learning. College Composition and Communication, 28(2), 122-128. https://doi.org/10.2307/356095

Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100(3), 363-406. https://doi.org/10.1037/0033-295X.100.3.363

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34(10), 906-911. https://doi.org/10.1037/0003-066X.34.10.906

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. College Composition and Communication, 32(4), 365-387. https://doi.org/10.2307/356600

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). How to design and evaluate research in education (8th ed.). McGraw-Hill.

Gupta, A., & Bhardwaj, A. (2023). Prompt engineering for generative AI in educational contexts: Challenges and opportunities. Journal of Educational Technology & Society, 26(3), 112-127.

Hake, R. R. (1999). Analyzing change/gain scores. American Educational Research Association, Division D. https://www.physics.indiana.edu/~sdi/AnalyzingChange-Gain.pdf

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112. https://doi.org/10.3102/003465430298487

Huang, J., Saleh, S., & Liu, Y. (2022). A review on artificial intelligence in education: Chatbots and their applications, benefits and challenges. Sciences of Education, 11(2), 78-94.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. Educational Psychologist, 38(1), 23-31. https://doi.org/10.1207/S15326985EP3801_4

King, A. (1991). Effects of training in strategic questioning on children's problem-solving performance. Journal of Educational Psychology, 83(3), 307-317. https://doi.org/10.1037/0022-0663.83.3.307

Kofinas, A. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. British Journal of Educational Technology. https://doi.org/10.1111/bjet.13585

Lawshe, C. H. (1975). A quantitative approach to content validity. Personnel Psychology, 28(4), 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence unleashed: An argument for AI in education. Pearson Education.

Lutfi, M. (2023). Metacognitive competencies in higher education: A systematic review of Indonesian-language teacher education programs. Jurnal Pendidikan Bahasa dan Sastra, 23(1), 45-62. https://doi.org/10.17509/bs_jpbsp.v23i1.56781

Sallam, M., Salim, N. A., Barakat, M., & Al-Tammemi, A. B. (2024). ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. Narra J, 3(1), e103. https://doi.org/10.52225/narra.v3i1.103

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. Contemporary Educational Psychology, 19(4), 460-475. https://doi.org/10.1006/ceps.1994.1033

Selwyn, N. (2016). Is technology good for education? Polity Press.

Setiawan, A., & Suwatno, S. (2024). Theory-practice gap in pre-service language teacher education: Evidence from Indonesian higher education. Indonesian Journal of Applied Linguistics, 13(3), 611-625. https://doi.org/10.17509/ijal.v13i3.62014

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. Harvard Educational Review, 57(1), 1-22. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24) (pp. 1-24). ACM. https://doi.org/10.1145/3613904.3642902

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. Metacognition and Learning, 1(1), 3-14. https://doi.org/10.1007/s11409-006-6893-0

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 13-39). Academic Press.

Zimmerman, B. J., & Schunk, D. H. (Eds.). (2019). Self-regulated learning and academic achievement: Theoretical perspectives (3rd ed.). Routledge.