

The Paradox of AI-Assisted Learning: Analyzing Self-Regulated Learning Readiness Toward the Authenticity of Coding and AI Learning Outcomes

Fatmalia Nur Prihastuti^{1,*}, Mukhammad Luqman Hakim¹, & Edward Harefa²

¹Department of Learning Technology, Universitas Negeri Yogyakarta, Indonesia

²School of Educational Technology, Beijing Normal University, China

*Corresponding email: fatmalianur.2025@student.uny.ac.id

Received: 05 June 2026

Accepted: 29 August 2026

Published: 22 September 2026

Abstract: This study examines how self-regulated learning (SRL) readiness relates to learning outcomes when students work with AI assistance versus when AI is withheld, in the KKA (*Komputer, Koding, dan Kecerdasan Buatan*) subject. A second aim is to propose and test the construct of pseudo self-regulated learning (pseudo-SRL) within prompt engineering tasks for AI-based image generation using Google Gemini. This quantitative correlational study involved 33 grade XI students at a senior high school based in a pesantren. Each student completed a 29-item SRL questionnaire, an AI-assisted practical test in which students created and revised prompts to generate and edit images, and a closed-book written test on prompt structure, AI limitations, and ethics, completed without AI support, as in a mid-semester test. Descriptive statistics, Pearson correlation, paired-samples t-tests, and one-way ANOVA were used to analyze the data. Students scored markedly higher on the practical task ($M = 78.8$) than on the written test ($M = 60.8$), producing a performance-knowledge gap of $t(32) = 7.63$, $p < .001$, 95% CI [13.2, 22.8]. SRL showed only a weak, non-significant relationship with written test performance ($r = .298$, $p = .092$) and with practical scores ($r = .125$, $p = .487$); practical and written scores were themselves significantly correlated ($r = .661$, $p < .001$). The gap was descriptively narrowest among students with high SRL (12.0 points), though a one-way ANOVA showed this difference across SRL tertiles was not statistically significant, $F(2,30) = 1.81$, $p = .181$. No participants met the operational criteria for pseudo-SRL, and an exploratory clustering analysis also found no natural pseudo-SRL profile, suggesting that other factors drove the performance-knowledge gap. The paradox of AI-assisted learning also affects prompt-engineering tasks in KKA. We refer to students who fail to translate their metacognitive self-reports into an autonomous understanding of prompting principles as pseudo-SRL; this gap is masked by their proficiency with AI tools. Assessment designs that combine behavior-based self-regulation measures with both AI-assisted and AI-free tasks might be valuable in programming and AI education. The results point to the need for KKA curriculum designers to incorporate dual-assessment models that differentiate between independent mastery of concepts and performance with AI assistance, to allow AI tools to augment rather than substitute for foundational knowledge.

Keywords: Generative AI, self-regulated learning, pseudo-SRL, programming education, performance-knowledge gap.

Article's DOI: <https://doi.org/10.23960/jpp.v16i3.pp1849-1863>

■ INTRODUCTION

The fast-paced development of Generative AI technology as a part of the everyday lives of people using digital devices is one of the factors contributing to the implementation of the KKA

subject in Indonesia's secondary education. The KKA course is no longer an addition to the curriculum; rather, it is the very basis of how computer science and programming should be learned. Furthermore, students can utilize iterative

prompts to create and edit images with the help of such technologies as Google Gemini and others; that is something regular e-learning systems and books can never offer. Once a student had to understand programming principles from a printed module, now they can test their assumptions with an AI tool. The KKA course consists of 3 pillars: computer science, programming logic, and AI literacy.

KKA's artificial intelligence focus is essentially prompt engineering, the art of sculpting and polishing natural-language instructions until a model such as Gemini, DALL-E, or Midjourney yields the desired result (Liu & Chilton, 2022). A student can create professional results in one take with a well-formed prompt, or they can take a shot and turn it into something else with many layers of editing." A growing body of research points to a troubling pattern behind the apparent ease of use of AI tools: students who seem quite capable when using the tools are often unable to explain why a prompt worked, what the model cannot do, or where the ethical boundaries are when the tools are not in use (Bassner et al., 2026; Wang et al., 2025). In a meta-analysis of 69 experimental studies, Deng et al. (2025) reported significant improvements in academic performance ($g+ = 0.712$) and motivation ($g+ = 0.881$) when applying ChatGPT. However, they also pointed to a methodological problem that pervades this literature. When AI is present at the time of the outcome measure, the "learning" measured might reflect the quality of the AI's output. In this study, we refer to the gap between a student's real understanding and their potential with AI as the "AI-assisted learning paradox".

One explanation for this apparent contradiction may be self-regulated learning (SRL). Self-regulated learning involves learners setting goals, monitoring their progress, seeking help when necessary, and adjusting their strategies based on what they learn (Zimmerman, 2000, 2002). An independent learner would use Gemini

as a support network, not a crutch, by trying several prompt versions, comparing the results, and then asking why one version was more effective than the others. Students are able to describe good learning strategies, but when the task becomes more difficult, they find it difficult to use them (Viberg et al., 2025). However, outsourcing cognitive tasks to AI might widen the divide rather than bridge it.

Much emphasis has been placed on the perils of over-reliance on GenAI. While GenAI facilitates interaction and makes creation more inclusive, Nathaniel et al. (2025) argue that these technologies may hinder the acquisition of critical knowledge and higher-order thinking when educators fail to use them strategically. Drawing on SRL theory, Universal Design for Learning, and Bloom's taxonomy, the researchers introduced the GenAI-Ped paradigm, which educators can use to balance AI assistance with skills acquisition. In contrast, effective picture-generation prompting requires specific knowledge such as syntax, style modifiers, aspect ratio, negative prompting, and iteration (Liu & Chilton, 2022). One of the explicitly stated goals of KKA's AI strand is instant literacy; however, students who use Gemini for picture generation without understanding why a particular prompt was used fail to achieve it.

In this paper, we introduce a notion of pseudo-self-regulated learning (pseudo-SRL). It describes a situation in which a student performs well on a validated SRL test, indicating strong metacognitive skills. However, removing the AI does not turn this ability into performance. Pseudo-SRL performs well on KKA's prompt engineering tasks, producing visually pleasing Gemini-generated graphics (high SRL, high practical score). However, she gets no points for writing reasoning. She does not illuminate the model's limitations. She does not explain why the given prompt was successful. She does not show how to write a successful prompt herself. The

result is a false confidence in the skill that misleads both teacher and student. In a coding and AI classroom, Pseudo-SRL offers an operational definition and an empirical test that draws on the issues of overreliance and assessment validity raised by Nathaniel et al. (2025) and Deng et al. (2025). Pseudo-SRL provides an operational definition and conducts an empirical test in a coding and AI classroom.

Another key point to consider is the study's context: an Islamic boarding school (Pesantren). In this kind of institution, students have to balance religious and academic education. Furthermore, the authors note that children receive very little parental supervision in their studying. Although Hamid and Sari (2022) do not mention the role of AI technologies in learning, they show that SRL predicts success in such circumstances. The relationship between SRL and AI tools in a pesantren is not clear-cut.

To address the research gap outlined above, this study was guided by three research questions, each accompanied by a corresponding hypothesis:

- RQ1: How does SRL readiness relate to student learning outcomes in KKA when AI assistance is available compared to when it is withheld?
- H1: SRL readiness will show a stronger positive correlation with written (AI-free) test scores than with practical (AI-assisted) scores.
- RQ2: To what extent does students' SRL level influence the magnitude of the performance-knowledge gap?
- H2: Students with higher SRL levels will exhibit a narrower performance-knowledge gap compared to those with lower SRL levels.
- RQ3: To what extent does the pseudo-SRL construct apply to students in AI-based image generation tasks?
- H3: No students meeting the strict pseudo-SRL criteria ($SRL \geq 120$, $Practicale \geq 80$, $Writtend \leq 45$) will be found in the sample,

suggesting that high self-reported SRL may be sufficient to prevent severe performance-knowledge gaps.

METHOD

Research Design

The study used a quantitative correlational design. The study used an AI-assisted practical test, in which students were guided to create and edit images using Google Gemini through detailed, iterative prompting, and a closed-book written test, in which students were asked to cover topics such as prompt structure, AI limitations, and ethical considerations, as the study's anchor assessments. Before both evaluations, SRL preparedness was gauged using a self-report questionnaire.

Participants

The sample was 33 eleventh-grade students (semester 2) taking the KKA (*Komputer, Koding, dan Kecerdasan Buatan*) course at an Indonesian pesantren-based senior high school. The inclusion criteria were: (1) active participation in the KKA course in the academic year 2025/2026, (2) successful completion of both the practical and written assessments, and (3) willingness to participate, as expressed in the informed consent. Students with prior formal training in programming other than the KKA curriculum were excluded for baseline homogeneity.

The use of generative AI tools (e.g., ChatGPT, Gemini, or similar) before the study was reported by 18 students (54.5%), mostly for school assignments and general information seeking. Nevertheless, none of the students had formal training in prompt engineering before the KKA course. The demographic profile of the sample is shown in Table 1.

Student identities have been anonymized to preserve privacy. In the supplementary material, they are only identified by their anonymous codes (S1-S33).

Table 1. Demographic characteristics of participants (N = 33)

Characteristic	Category	n	%
Gender	Male	15	45.5
	Female	18	54.5
Grade Level	11th Grade	33	100
Semester	2 (Even)	33	100
Prior AI Tool Experience	Yes	18	54.5
	No	15	45.5
Prior Prompt Engineering Training	Yes	0	0
	No	33	100

Sample Size Justification

One notable limitation of the experimental design is that the sample size was set at 33, reflecting the number of students in a single KKA class in pesantren schools. A post-hoc power analysis was conducted using G*Power 3.1 (Faul et al., 2009) to assess whether the sample size remained sufficient for significant inference. Using $N=33$, a medium effect size ($r=.30$), and $\alpha=.05$, the primary correlation analysis yields an achieved power of approximately .42, which falls well below the conventional .80 benchmark but is still considered feasible for exploratory work in education research, according to Cohen (1992). The one-way ANOVA with the three 11-groups performs worse: the attained power is just .48 at a medium effect size ($f=.25$), indicating that the design is really weak for small-to-medium effects. The associations shown here were nonetheless discernible despite the small N since our actual observations, with $\eta^2 = .108$ for the SRL-tertile ANOVA and $r = .298$ for the SRL-written correlation, are modest and did not reach statistical significance at this sample size. We point this out as a serious drawback and stress the need for large-scale replication. To give you an idea of the scope of the problem, similar studies investigating AI-assisted learning in niche educational contexts have used sample sizes of 30–40 students (e.g., Kazemitabaar et al., 2023; Yilmaz & Yilmaz, 2023).

Instrument Validation

The practical test was administered in one 60-minute session in the school's computer laboratory. Every student was provided with a personal computer with internet access and a Google Gemini account. The process was:

1. The researcher gave a 10-minute orientation to the task objective, rubric criteria, and an example of an effective prompt. Students were asked to produce an image of a 'traditional Indonesian market scene' with several visual characteristics.
2. Students were allotted 5 minutes to design their first prompt that should contain: (a) a clear description of the subject, (b) at least three style or composition modifiers (e.g., watercolor, lighting, perspective), and (c) a negative prompt to exclude undesired elements.
3. Students were given 30 minutes to write their first prompt, generate the image with Google Gemini, and iteratively improve their prompt based on AI's output. A maximum of five cycles of revision was allowed.
4. A worksheet was provided to the students during the refinement process for them to keep a record of each version of the prompt and the image output produced.
5. The last 15 minutes were given for students to select their best image and submit the final prompt and the image file.

The specific instruction given to students was: 'Create a prompt to generate an image of a traditional market in Indonesia. Your prompt must include details about the setting, objects, people, atmosphere, style, lighting, and composition. Use negative prompts to avoid unwanted elements. You may revise your prompt based on the AI output. On the walls are hanging hooks, pots, and pans. There is a table in the middle of the room, a couple of chairs, and a window over the sink. The lighting is dark and moody, with just one light bulb hanging from the ceiling. The composition is strict and claustrophobic, emphasizing the tight space. The style is realistic with a focus on detail and texture. No people are present. Negative prompts remove unwanted elements. You are able to change your prompt based on the AI's reply.

Inter-Rater Reliability

The practical-test answer of each participant was scored independently by two raters, including the main researcher and a KKA teacher with more than 5 years of experience in teaching programming. Both raters were trained on the rubric using a separate pilot set of responses, reviewing the four rubric components (prompt clarity, effectiveness, final image quality, efficiency) with anchor examples; they then independently scored ten practice responses and discussed any disagreements to calibrate their judgments before the main assessment. Agreement was evaluated using Cohen's weighted kappa for each rubric component and the intraclass correlation coefficient (ICC) for total scores. Weighted kappa for the four components ranged from .78 to .89, indicating substantial to almost perfect agreement according to Landis and Koch (1977) benchmarks. The ICC for total practical scores was .91 (95% CI [.84, .95]), indicating excellent reliability. All disagreements were resolved by consensus, and the analysis used the average of the two raters' scores. The full rubric, sample

responses, and rater training materials are provided in the supplementary materials.

Instruments

SRL Questionnaire: A 29-item, 5-point Likert instrument (1 = strongly disagree, 5 = strongly agree), adapted from the Motivated Strategies for Learning Questionnaire and Zimmerman's SRLIS, covered six dimensions: goal setting and planning, self-monitoring, environment structuring, help-seeking, self-evaluation, and motivation. Internal consistency was acceptable (Cronbach's $\alpha = .84$).

To establish construct validity for the pesantren context, item-total correlation analysis was performed directly on the main sample data ($N = 33$). An item was considered valid if its correlation coefficient exceeded the critical r-table value of 0.296 ($df = 31$, $\alpha = .05$, one-tailed). The analysis revealed that item-total correlation coefficients ranged from 0.324 to 0.761 across the 29 items. Two items (items 8 and 17) had correlation coefficients below the critical threshold ($r = .289$ and $.275$, respectively) and were removed, resulting in a 27-item final version. The remaining 27 items all demonstrated significant correlations with the total score ($p < .05$), confirming construct validity. The revised instrument maintained a high internal consistency (Cronbach's $\alpha = .87$).

Practical Prompt Engineering Test (With AI Assistance): In this practical assessment, students completed one task aligned with the KKA learning objectives: writing an elaborate prompt to generate an image using Google Gemini and then enhancing the provided image using prompt engineering techniques. This assignment helped in assessing: (1) How clear and specific the prompts were in terms of the subject, style, lighting, composition, and negative prompts; (2) How well the prompts were adjusted according to the output from the AI; and (3) The quality of the final image.

Written Test (AI-prohibited): A 60-minute, closed-book test made up of scenario-based questions (reproduced in the supplementary materials) probed understanding of prompt engineering without any AI or internet access. Items asked students to judge whether AI can infer fruit sweetness from an image and why; explain how AI reads image features such as color, texture, and size; identify effective prompt structures; name factors behind inaccurate predictions; describe how data quality shapes AI analysis; propose human verification strategies; and weigh AI's proper role as a decision support tool rather than a final decision-maker. Scoring (0–100) followed a validated rubric.

Operational Definition of Pseudo-SRL

Three cut-offs, set a priori from the score distribution and from theory, define pseudo-SRL operationally:

1. SRL total ≥ 120 : This value marks the top tertile (67th percentile) of the SRL distribution ($M = 114.2$, $SD = 11.0$), roughly 0.53 SD above the mean, capturing students who report the strongest self-regulatory capacity: goal setting, self-monitoring, and strategic planning. By conventional SRL theory, these are exactly the students expected to show strong, independent learning outcomes.
2. Practical score ≥ 80 : Rounded up from the sample mean of 78.8, this represents roughly 0.07 SD above the mean and the upper half of the score distribution. The threshold is deliberately modest: it confirms genuine skill with Gemini-based image generation without implying that pseudo-SRL students are weak learners overall; their strength is specific to the AI-assisted condition.
3. Written test score ≤ 45 : Set one SD (16.1) below the sample mean of 60.8 ($60.8 - 16.1 = 44.7 \approx 45$); this falls near the 16th percentile of independent conceptual knowledge under

a normal distribution. A score at or below 45 means a student cannot demonstrate basic command of prompt engineering, AI limitations, or verification strategies once AI support is gone, well below the 65–70 threshold normally used as a passing grade in the Indonesian system.

Theoretical Rationale

These three criteria were chosen together, a priori, to separate three constructs that are otherwise easy to conflate: metacognitive awareness (self-reported SRL), AI tool proficiency (practical score), and conceptual understanding (written score). A pseudo-SRL classification flags students who score high on the first two but low on the third, a pattern consistent with genuine overreliance on AI despite apparent self-regulatory strength. The idea builds on the “overreliance without understanding” that (Nathaniel et al., 2025) describe and the “assessment validity confounds” that Deng et al. (2025) raise. We note that these specific cut-offs are tied to this sample and setting; future studies should test and possibly refine them using ROC analysis or other criterion-referenced methods.

The threshold of ≤ 45 for the written test was established through multiple pedagogical and empirical considerations. First, within the Indonesian secondary education system, the minimum mastery criterion (Kriteria Ketuntasan Minimal/KKM) for cognitive subjects typically ranges between 65 and 75 (Kemendikbud, 2016). A score of 45 falls substantially below this national benchmark, signifying a clear failure to achieve even partial mastery of the assessed competencies. Second, from a criterion-referenced measurement perspective, a score below half of the total possible points indicates that a student cannot correctly answer the majority of fundamental items related to prompt structure, AI limitations, and verification strategies, thereby operationalizing a state of “non-

competence” (Nitko & Brookhart, 2014); a cut-off of 45 falls comfortably within this non-competence band. Third, empirically, the score of 45 is precisely one standard deviation (16.1) below the sample mean ($60.8 - 16.1 = 44.7 \approx 45$), placing students at approximately the 16th percentile under a normal distribution a standard statistical cut-off for identifying “low-performing” or “at-risk” subgroups in educational research (Field, 2018; Cohen, 1992). Fourth, this threshold aligns with the local school’s grading policy, where scores below 60 are categorized as “insufficient” or “failing” for basic competency mastery. Therefore, the d’ 45 cut-off is neither arbitrary nor sample-dependent; it represents a convergence of national norms, measurement theory, empirical distribution, and institutional standards.

Alternative Approaches Considered

The standardized residuals (real versus predicted written scores), cluster analysis (which permits the formation of sub-categories in the data), and discrepancy scores (practical versus written) compared with a threshold were the three options before we selected fixed a priori cut-offs. The categorizations resulting from residual and cluster approaches may be too sample-dependent to replicate, and all three require a larger sample size than we achieved. Although the fixed a priori cut-off is entirely subjective, it allows others to replicate the analysis.

Procedure

The SRL questionnaire was sent online one week before the assessments. The practical and written tests were administered on two separate days, in counterbalanced order, with approval

from the institutional ethics committee covering all procedures.

Data Analysis

Descriptive statistics, Pearson correlations, and paired t-tests formed the core analysis, with students split into SRL tertiles (high, medium, low) for group comparisons. Pseudo-SRL followed the three a priori cut-offs already described: SRL total ≥ 120 (top tertile), practical score ≥ 80 (above the sample mean of 78.8), and written score ≤ 45 (one SD below the sample mean of 60.8) together isolating students whose metacognitive self-report and AI-assisted performance outpace their independent conceptual grasp. Although these criteria were established a priori, their application to the current sample did not yield any participants meeting all three thresholds simultaneously.

Parametric assumptions were checked before running the main analyses. Shapiro–Wilk tests showed no meaningful departure from normality for SRL total ($W = .967, p = .412$), practical scores ($W = .945, p = .098$), or written scores ($W = .981, p = .842$). Levene’s test likewise supported homogeneity of variance for the one-way ANOVA for both practical scores ($F(2,30) = 0.74, p = .485$) and written scores ($F(2,30) = 1.62, p = .214$).

■ RESULT AND DISCUSSION

Descriptive Statistics and the Performance–Knowledge Gap

Table 1 lays out descriptive statistics for SRL total scores, practical prompt engineering scores (AI-assisted image generation and editing), and written test scores (prompt engineering knowledge without AI) across the full sample.

Table 1. Descriptive statistics of SRL, practical, and written test scores ($N = 33$)

Variable	Mean	SD	Minimum	Maximum	Skewness
SRL Total (max 145)	114.2	11.0	95	133	-0.26
Practical Prompt Engineering Score (AI-assisted)	78.8	16.7	40	97	-0.52
Written Test Score (AI-prohibited)	60.8	16.1	30	94	-0.03

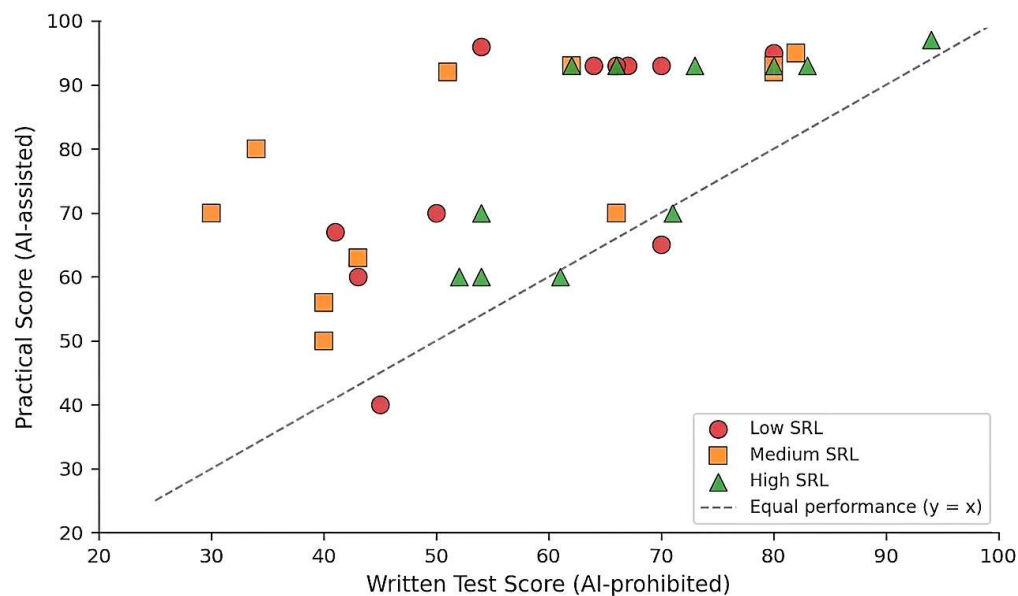


Figure 1. Scatter plot of individual practical (AI-assisted) and written (AI-prohibited) test scores, color- and marker-coded by SRL tertile (N = 33). The diagonal line ($y = x$) represents equal performance; points above the line indicate a positive performance–knowledge gap (practical > written)

The difference between mean written and practical exam scores was significant, at 18.0 points (paired $t(32) = 7.63$, $p < .001$, 95% CI [13.2, 22.8]). The knowledge-practice discrepancy showed a positive difference in practical exam scores over written exam scores in 29 of 33 subjects (87.9%), excellent replication of the AI-based learning paradox within KKA. Only four students (12.1%) scored higher on their written exams than on their practical exams.

Results for RQ1/H1: Correlations Between SRL, Practical, and Written Scores

As shown in Figure 1 above, individual performance-knowledge gaps varied considerably across students. To test H1, Pearson correlations were computed between SRL total scores and each performance measure, as well as between the two performance measures themselves. Table 2 reports the results in APA Style.

Table 3. Pearson correlations among SRL total, practical score, and written test score (N = 33)

Variable Pair	r	p	N
SRL Total — Written Test Score	.298	.092	33
SRL Total — Practical Score	.125	.487	33
Practical Score — Written Score	.661	<.001	33

Contrary to H1, SRL total did not correlate significantly with either written or practical scores; both associations were weak. Practical and written scores, however, were significantly and moderately to strongly correlated, indicating that students who performed better with AI assistance also tended to perform better independently; the two are related. However, a large and systematic gap remains between them regardless of SRL readiness.

Results for RQ2/H2: SRL Level and the Magnitude of the Performance–Knowledge Gap

To test H2, students were split into SRL tertiles (high, medium, low), and the practical-written gap was compared across groups. Descriptively, the gap was narrowest among students in the high SRL tertile (12.0 points); the medium (22.4 points) and low (19.5 points) tertiles did not follow a clean monotonic pattern.

A one-way ANOVA confirmed that the gap did not differ significantly across SRL tertiles, $F(2,30) = 1.81$, $p = .181$, $\eta^2 = .108$, and a supplementary linear regression of the gap on continuous SRL scores was also not significant, $b = -0.25$, $p = .265$, $R^2 = .040$. H2 is therefore not statistically supported, although the descriptive trend for the high-SRL group (Figure 2) leaned in the predicted direction.

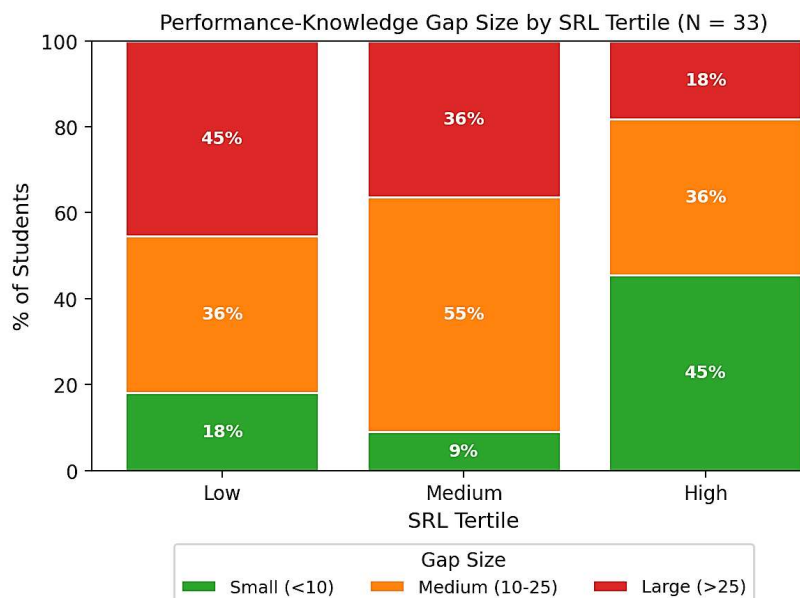


Figure 2. Performance-knowledge gap size (small/medium/large) as a percentage of students within each SRL tertile (N = 33)

Results for RQ3/H3: Prevalence of the Pseudo-SRL Construct

To test H3, students were screened against the three a priori pseudo-SRL criteria (SRL total ≥ 120 , practical score ≥ 80 , written score ≤ 45). No participants met all three thresholds; therefore, no pseudo-SRL cases were identified. This finding is consistent with H3.

As a data-driven complement to the a priori pseudo-SRL cut-offs, an exploratory K-means cluster analysis ($k = 3$, standardized SRL, practical, and written scores; silhouette = .35) tested whether a “high-SRL, high-practical, low-written” profile would emerge naturally from the data. It did not: the three clusters that emerged

were a high-practical/moderate-written group with below-average SRL ($n = 10$; $M_{\text{SRL}} = 103.8$, $M_{\text{practical}} = 86.0$, $M_{\text{written}} = 63.8$), a low-performing group across both measures ($n = 12$; $M_{\text{SRL}} = 114.8$, $M_{\text{practical}} = 61.3$, $M_{\text{written}} = 44.8$), and a high-achieving group scoring above average on all three measures, including written performance ($n = 11$; $M_{\text{SRL}} = 123.0$, $M_{\text{practical}} = 91.4$, $M_{\text{written}} = 75.7$). None of the three data-driven profiles combined high SRL and high practical performance with low written performance, converging with the a priori finding of zero pseudo-SRL cases (H3) through an independent, distribution-based method.

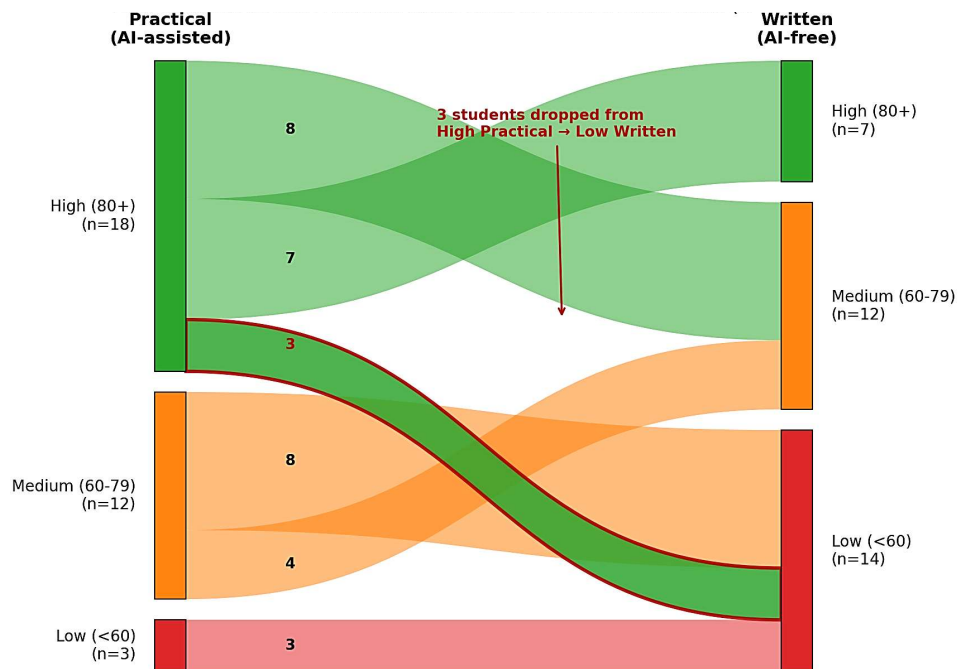


Figure 3. Flow of students from practical (AI-assisted) score band to written (AI-free) score band (N=33). The highlighted flow shows students who scored High on the practical task but Low on the written test

The results recreate the AI-assisted learning paradox in a novel and previously unexplored scenario: prompt engineering for AI picture production in KKA. The difference between performance and knowledge is 18.0 points, and the connection between practical and written scores is significant and moderate-to-strong ($r = .661$, $p < .001$). Despite this significant correlation, AI-assisted performance still substantially overestimates students' independent comprehension of the model's capabilities and limitations, as well as their ability to produce good prompts without assistance.

On another front, our statistics agree with Nathaniel et al. (2025), who claim that GenAI integration is most effective when training is intentional, and evaluation is systematic. In this scenario, SRL readiness did not significantly buffer the difference: descriptively, it was 12.0 points for students with good SRL preparation and 19.5 points for those with low SRL readiness (22.4 points for medium SRL), but a one-way ANOVA

showed this pattern was not statistically significant, $F(2,30) = 1.81$, $p = .181$. Furthermore, SRL did not significantly predict students' written results ($r = .298$, $p = .092$) or their practical scores ($r = .125$, $p = .487$), which does not support the hypothesis that AI technologies can level the playing field for students with poor self-regulation.

This finding aligns with a growing body of international research documenting the dissociation between AI-assisted performance and genuine conceptual understanding. In a large-scale randomized controlled trial involving 275 introductory programming students, Bassner et al. (2026) found that while both scaffolded and unrestricted AI support significantly improved exercise scores compared to a no-AI control group, neither condition produced greater pre-post knowledge gains or code-comprehension advantages leading the authors to conclude that generative AI acted "primarily as a performance aid rather than a learning enhancer". Similarly, Viberg et al. (2025) analyzed 364 student prompts

in programming courses and found that students predominantly relied on LLMs for elaboration and help-seeking, while self-regulated learning strategies such as effort regulation, critical thinking, and organization were used far less frequently. This pattern of passive reliance resonates with our observation that students achieved high practical scores with Gemini yet struggled to articulate prompt structures or AI limitations without assistance. However, the literature is not univocal. Some studies have reported more optimistic outcomes: Li et al. (2025) found that generative AI-supported programming learning significantly improved learning outcomes, motivation, and self-efficacy among students with certain learning styles, while others have documented mixed gains in conceptual understanding—lower for some topics but higher for others (e.g., functions versus structures) (Andleeb et al., 2025). Moreover, research by Bouamor et al. (2025) suggests that while GenAI tools help less technically prepared students complete programming tasks, those who rely heavily on AI assistance struggle with more complex, transformative tasks such as closed-book exams. This finding directly parallels our own performance-knowledge gap. These contrasting findings underscore that the educational value of GenAI is not intrinsic to the technology itself but depends critically on instructional design, assessment strategies, and the development of students' self-regulatory capacities (Nathaniel et al., 2025).

Although this sample identified no pseudo-SRL cases, the substantial performance-knowledge gap across all SRL levels still requires three key considerations for KKA and AI-based education. Firstly, environments augmented with AI cannot be adequately identified through self-report SRL surveys; one can remain a pseudo-SRL, even with good results. To compensate for this deficiency, dual assessment is necessary—a practical assignment using AI and a conceptual one without AI. Second, SRL training should

include both fast-design training and conceptual training on how AI functions, its limitations, and how to validate its results. Thirdly, it is essential to address this matter in pesantren environments, where students already face the challenge of combining religious and academic activities (Hamid & Sari, 2022). An AI-assisted system for managing independent study would help in this case.

Limitations

Several limitations of this study are acknowledged. The sample size is modest ($N=33$), and the sample is collected from one pesantren-based institution, limiting the generalizability of the findings. Because this is a correlational study, the direction of causality between SRL and AI-assisted outcomes cannot be determined. SRL was measured only via self-report, which is susceptible to social desirability bias and may not reflect actual regulating behavior (Zimmerman, 2000). A no-AI control group was not included; thus, absolute learning gains with versus without AI cannot be directly compared. The cut-offs for the pseudo-SRL construct ($SRL \geq 120$, practical ≥ 80 , written ≤ 45) are by design arbitrary and may require tuning in future work. Relatedly, the study was underpowered to detect the hypothesized SRL-buffering effect: the one-way ANOVA on the gap by SRL tertile did not reach significance ($F(2,30) = 1.81$, $p = .181$, $\eta^2 = .108$), so H1 and H2 should be re-tested in a larger sample before being ruled out. However, the broader area still lacks defined frameworks for investigating GenAI-integrated learning, which Nathaniel et al. (2025) address. Lastly, the practical exam was a one-time prompting session and may not reflect a student's prompt-engineering capabilities.

Future Research

Longitudinal work could trace how pseudo-SRL develops as students spend more time with image-generation tools, while experimental

designs that randomly assign students to unrestricted AI, scaffolded prompting, or no-AI conditions would help establish causality and address the power and baseline concerns Deng et al. (2025) raise. Ethnographic case studies inside pesantren settings could also clarify how local cultural and religious norms interact with AI-assisted image creation and SRL. Beyond that, replication with larger and more varied samples different ages, different subjects beyond KKA remains a clear next step, alongside behavioral SRL instruments that read pseudo-SRL directly off prompt-iteration logs: number of revisions, variety of parameter changes, time spent reflecting between attempts. Testing and extending the GenAI-Ped framework (Nathaniel et al., 2025) to include pseudo-SRL detection would be a natural continuation of this line of work.

■ CONCLUSION

This study confirms an AI-assisted learning paradox within the KKA subject: students' performance on AI-assisted tasks substantially exceeded their independent conceptual understanding, and this performance-knowledge gap was statistically significant and present in the large majority of students.

Self-regulated learning readiness did not significantly buffer this gap in the present sample. Students with high SRL readiness showed the narrowest gap descriptively, but this pattern did not reach statistical significance across SRL tertiles. SRL was not significantly related to either written (AI-free) or AI-assisted practical performance, while practical and written performance were themselves significantly correlated, suggesting that AI assistance inflated performance fairly consistently across the ability range rather than in an SRL-dependent way.

Although the strict pseudo-SRL criteria yielded no cases in this sample, the conceptual framework remains valuable for understanding how high self-reported metacognition does not

automatically translate into independent mastery once AI support is removed. For KKA and AI education more broadly, these findings point to the need for dual-assessment designs that combine AI-assisted and AI-free tasks. Relying solely on AI-supported performance risks overestimating students' independent competencies; explicit instruction on AI's mechanisms and limitations, together with behavior-based markers of self-regulation, is needed so that AI functions as a support for deeper learning rather than a substitute for it.

■ DECLARATION OF GENERATIVE AI USAGE IN THE WRITING PROCESS

While writing this manuscript, the authors used Cloud AI, DeepSeek, and Google Gemini to create first drafts and language outlines. Every draft created was critically examined, cross-checked for veracity from primary sources, and extensively revised to create a final draft that accurately reflected the research's findings. The author is solely responsible for the final manuscript.

■ REFERENCES

- Achuthan, K. (2025). Artificial intelligence and learner autonomy: A meta-analysis of self-regulated and self-directed learning. *Frontiers in Education*, 10, Article 1738751. <https://doi.org/10.3389/feduc.2025.1738751>
- Andleeb, S., Zhao, J., Zhang, B., & He, Y. (2025). ChatGPT in introductory programming: Counterbalanced evaluation of code quality, conceptual learning, and student perceptions. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2501.12345>
- Bassner, P., Lenk-Ostendorf, B., Beinstingel, R., Wasner, T., & Krusche, S. (2026). Less stress, better scores, same learning: The

- dissociation of performance and learning in AI-supported programming education. *Computers and Education: Artificial Intelligence*, 10, 100537. <https://doi.org/10.1016/j.caeai.2025.100537>
- Bouamor, H., Gongora-Svartzman, G., Heimann, L., & Huang, S. (2025). Evaluating GenAI's effectiveness for students with varied programming backgrounds in a software development course. In *Proceedings of the 2025 ACM Conference* (pp. 1–8). ACM. <https://doi.org/10.1145/3641555.3705175>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, 227, 105224. <https://doi.org/10.1016/j.compedu.2024.105224>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Hamid, A., & Sari, D. P. (2022). Self-regulated learning and academic achievement in pesantren-based schools. *Jurnal Pendidikan Islam*, 8(2), 145–158. <https://doi.org/10.15575/jpi.v8i2.12345>
- Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11), 1539–1558. <https://doi.org/10.1002/sim.1186>
- Kazemitabaar, M., Chow, J., Ma, C. K. T., Ericson, B. J., Weintrop, D., & Grossman, T. (2023). Studying the effect of AI code generators on supporting novice learners in introductory programming. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–23). ACM. <https://doi.org/10.1145/3544548.3580919>
- Kong, S. C., & Yang, Y. (2024). A human-centered learning and teaching framework using generative artificial intelligence for self-regulated learning development through domain knowledge learning in K-12 settings. *IEEE Transactions on Learning Technologies*, 17, 1562–1573. <https://doi.org/10.1109/TLT.2024.3392830>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Li, S., Liu, J., & Dong, Q. (2025). Generative artificial intelligence-supported programming education: Effects on learning performance, self-efficacy, and processes. *Australasian Journal of Educational Technology*, 41(3), 1–25. <https://doi.org/10.14742/ajet.9932>
- Li, W., Huang, J. Y., Liu, C. Y., Tseng, J. C., & Wang, S. P. (2023). A study on the relationship between student learning engagements and higher-order thinking skills in programming learning. *Thinking Skills and Creativity*, 49, 101369. <https://doi.org/10.1016/j.tsc.2023.101369>
- Liu, V., & Chilton, L. B. (2022). Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. <https://doi.org/10.1145/3491102.3501825>
- Nathaniel, J., Oyelere, S. S., Suhonen, J., & Tedre, M. (2025). Literature review on the

- integration of generative AI in programming education. *International Journal of Artificial Intelligence in Education*, 35, 2724–2755. <https://doi.org/10.1007/s40593-025-00524-3>
- Nieto-Cardenas, J., Kramer, E. J., Kurto, P., Dickey, E., & Bejarano, A. (2026). OWLgorithm: Supporting self-regulated learning in competitive programming through LLM-driven reflection. In *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V.1* (SIGCSE TS, 2026) (pp. 1–7). ACM. <https://doi.org/10.1145/3770762.3772662>
- Oppenlaender, J. (2022). A taxonomy of prompt modifiers for text-to-image generation. *arXiv preprint arXiv:2204.13988*.
- Phung, T., Pádurean, V. A., Cambronero, J., Gulwani, S., Kohn, T., Majumdar, R., & Soares, G. (2023). Generative AI for programming education: Benchmarking ChatGPT, GPT-4, and human tutors. In *Proceedings of the 2023 ACM Conference on International Computing Education Research* (pp. 41–42). ACM. <https://doi.org/10.1145/3568812.3603476>
- Popat, S., & Starkey, L. (2019). Learning to code or coding to learn? A systematic review. *Computers & Education*, 128, 365–376. <https://doi.org/10.1016/j.compedu.2018.10.005>
- Prasad, P., & Sane, A. (2024). A self-regulated learning framework using generative AI and its application in CS educational intervention design. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V.1* (pp. 1070–1076). ACM. <https://doi.org/10.1145/3626252.3630828>
- Prather, J., Denny, P., Leinonen, J., Becker, B. A., Albluwi, I., Craig, M., Keuning, H., Kiesler, N., Kohn, T., Luxton-Reilly, A., MacNeil, S., Petersen, A., Pettit, R., Reeves, B. N., & Savelka, J. (2023). The robots are here: Navigating the generative AI revolution in computing education. In *Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education* (pp. 108–159). ACM. <https://doi.org/10.1145/3623762.3633499>
- Qian, K., Liu, S., Li, T., Rakoviæ, M., Li, X., Guan, R., et al. (2026). Towards reliable generative AI-driven scaffolding: Reducing hallucinations and enhancing quality in self-regulated learning support. *Computers & Education*, 240, 105448. <https://doi.org/10.1016/j.compedu.2025.105448>
- Ren, L., Lee, K., & May, L. (2025). A systematic review exploring AI's role in self-regulated learning within education contexts. *IEEE Access*, 13, 109771–109782. <https://doi.org/10.1109/ACCESS.2025.3582600>
- Scherer, R., Siddiq, F., & Viveros, B. S. (2020). A meta-analysis of teaching and learning computer programming: Effective instructional approaches and conditions. *Computers in Human Behavior*, 109, 106349. <https://doi.org/10.1016/j.chb.2020.106349>
- Viberg, O., Wong, J., Feldman-Maggor, Y., Dunder, N., & Epp, C. D. (2025). Chatting with code: Exploring LLMs as learning partners in programming education. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), **Artificial Intelligence in Education - 26th International Conference, AIED 2025, Proceedings** (pp. 453–461). Springer.
- Wang, Z., Zou, D., Zhang, R., Lee, L.-K., Xie, H., & Wang, F. L. (2025). ChatGPT-enhanced self-regulated learning in

- programming education: Impacts on motivation, self-efficacy, and learning outcomes. Interactive Learning Environments. *Advance online publication*. <https://doi.org/10.1080/10494820.2025.2559919>
- Weng, X., Xia, Q., Ahmad, Z., & Chiu, T. K. F. (2024). Personality traits for self-regulated learning with generative artificial intelligence: The case of ChatGPT. *Computers and Education: Artificial Intelligence*, 7, 100315.
- Yilmaz, R., & Yilmaz, F. G. K. (2023). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy, and motivation. *Computers and Education: Artificial Intelligence*, 4, 100147. <https://doi.org/10.1016/j.caeai.2023.100147>
- Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11, 28.
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press.
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2