**Reforma: Jurnal Pendidikan dan Pembelajaran**
**Faculty of Teacher Training and Education**
**Universitas Islam Lamongan**
https://jurnalpendidikan.unisla.ac.id/index.php/reforma

*May 2026 Vol. 16, No. 1*
*e-ISSN*: 2621-4172
*p-ISSN*: 2503-1228
*pp. 30-47*

# Metacognitive Scaffolding through AI-Assisted Digital Annotation: Developing Argument Evaluation and Inference-Making in EFL Academic Reading

***Yanti Anggraini**

Department of Law, Faculty of Law, Universitas Dharma Indonesia, Universitas Dharma Indonesia. Serang Highway No. 18, Sukanegara, Cikupa District, Tangerang Regency, Banten 15710

*Corresponding e-mail: yeongi09@yahoo.com

Received: 04 May 2026; Revised: 26 May 2026; Accepted: 30 May 2026; Published: 31 May 2026

**Abstract**

*The integration of generative AI into academic reading contexts presents both a transformative scaffolding opportunity and a pedagogical risk: without deliberate instructional design, students may default to passive, query-and-paste AI use that reinforces surface-level text processing rather than analytical engagement. This Classroom Action Research (CAR) study investigated the effectiveness of AI-assisted digital annotation as a pedagogical intervention for improving undergraduate EFL students' critical reading of academic texts. Guided by the Kemmis and McTaggart spiral model, the study was implemented across two iterative cycles with a class of 50 undergraduate students enrolled in Bahasa Inggris course at an Indonesian higher education institution. Data were collected through Critical Reading Performance Tests (pre-test and post-test per cycle) and classroom observation, and were analysed using a mixed-method approach integrating paired-samples t-tests with thematic analysis following Miles, Huberman, and Saldaña's interactive model. Findings demonstrated statistically significant improvements in critical reading performance across both cycles, with a cumulative mean gain of 28.8 points from pre-test ($M = 49.8$) to the Cycle II post-test ($M = 78.6$; $p < .001$, Cohen's $d = 1.59$), and 83.3% of students surpassing the predetermined success threshold. Qualitative findings documented a shift from passive paraphrastic annotation to interrogative critical dialogue with both text and AI tool. The study concluded that AI-assisted digital annotation, when embedded within a structured pedagogical framework and accompanied by explicit instruction in critical AI interrogation, constituted an effective and ecologically valid strategy for developing students' critical reading competencies in EFL higher education contexts.*

**Keywords**: AI-assisted learning; critical reading; EFL; Classroom Action Research; metacognitive awareness.

**How to Cite:** Anggraini, Y. (2026). Metacognitive scaffolding through AI-assisted digital annotation: Developing argument evaluation and inference-making in EFL academic reading. *Reforma: Jurnal Pendidikan dan Pembelajaran*, 16(1), 30–47. https://doi.org/10.30736/reforma.v16i1.1369

https://doi.org/10.30736/reforma.v16i1.1369

Copyright©Anggraini, Y.
This is an open-access article under the CC-BY-SA License.

## INTRODUCTION

The ability to read academic texts critically is widely recognised as a foundational competency for success in higher education. Critical reading extends well beyond surface-level comprehension, requiring learners to analyse, evaluate,

question, and synthesise written arguments to construct informed judgements about the knowledge claims advanced by academic authors (Du et al., 2024). In the context of English as a Foreign Language (EFL) higher education, where students must simultaneously navigate linguistic complexity and academic discourse conventions, this competency assumes even greater salience. The growing presence of digital technologies in educational environments has transformed how students interact with texts, introducing both new affordances and new challenges for reading instruction. Digital annotation tools—platforms that allow learners to highlight, comment on, question, and respond to texts in digital formats—have emerged as a pedagogically promising strategy for fostering active, critical engagement with academic texts, with research confirming their positive impact on reading comprehension, metacognitive awareness, and critical thinking abilities (Li et al., 2016). More recently, the rapid proliferation of artificial intelligence (AI) tools in educational settings has opened further possibilities for personalised, adaptive, and scaffolded reading support. AI-assisted features embedded within annotation platforms—such as intelligent prompts, contextual clarification, and formative feedback on analytical comments—can serve as cognitive scaffolds that mediate students' engagement with complex academic texts, potentially transforming AI from a passive information retrieval tool into an active partner in the development of critical literacy (Fu & Hiniker, 2025). The convergence of digital annotation and AI assistance therefore represents a particularly fertile pedagogical intersection for advancing students' academic reading competencies. The intervention is theoretically grounded in Vygotsky (1978) Zone of Proximal Development (ZPD), which posits that learners achieve more complex cognitive tasks with the mediation of a more capable other—whether human or, in this case, an AI-assisted annotation system—than they could accomplish independently. Complementarily, cognitive load theory (Sweller, 1988) provides a framework for understanding how AI assistance reduces extraneous cognitive load, freeing attentional resources for higher-order analytical engagement with text.

Despite the acknowledged importance of critical reading in higher education, a well-documented and persistent gap exists between the reading competencies that academic study demands and those that students in EFL contexts actually possess. Needs analysis studies consistently identify higher-order critical reading skills—specifically argument evaluation, inference-making, and ideational synthesis—as the weakest areas among EFL university students, even as both students and teachers affirm their central importance (Aziz & Rawian, 2022). This deficiency is exacerbated by structural pedagogical limitations: research reveals a pervasive lack of initiative among EFL educators to integrate explicit critical reading instruction within their pedagogy, meaning that students are frequently expected to develop analytical reading competencies without the deliberate instructional scaffolding needed to do so (Du et al., 2024). The integration of AI tools into academic reading contexts, while offering genuine cognitive benefits, has simultaneously introduced a new pedagogical concern: students tend toward uncritical acceptance of AI-generated outputs, using AI tools in a passive, query-and-paste manner that may reinforce surface-level text processing rather than promoting deeper analytical engagement (Sublime & Renna, 2024). Taken together, these challenges—the persistent underdevelopment of critical reading skills, the inadequacy of existing instructional approaches, and the risk of

uncritical AI dependency—constitute a significant pedagogical problem that demands evidence-based, innovative instructional intervention. The challenge is compounded in the Indonesian higher education context, where EFL learners face particular difficulty in engaging analytically with the argumentative structures and inferential demands of English-medium academic texts, yet receive limited explicit instruction in how to do so. This challenge is not unique to Indonesia but reflects a systemic global pattern documented across EFL higher education contexts in Asia (Kusumawardana & Akhiriyah, 2022; Le & Nguyen, 2024), Africa (Gebremariam & Weldeyohannes, 2025; Tesfay & Weldeyohannes, 2024), and the Middle East (Al-Zahrani & Alasmari, 2025), where EFL learners face structurally comparable challenges in developing higher-order analytical reading competencies within English-medium academic discourse. The Indonesian context serves in this study as an illustrative and theoretically generative site for examining a problem of international pedagogical significance.

A growing body of research has investigated digital annotation and AI-assisted learning as separate pedagogical phenomena. Studies on digital annotation confirm its effectiveness in promoting active reading behaviours, collaborative engagement, and metacognitive awareness in L2 reading contexts (Velásquez, 2025). Research on AI in academic literacy demonstrates its potential to improve reading comprehension, support inferencing, and provide adaptive scaffolding, while also documenting the risks of overreliance and diminished metacognitive engagement (Fan, 2025). Similarly, Classroom Action Research (CAR) has established itself as a credible and ecologically valid methodology for investigating and improving context-specific pedagogical problems, with its iterative, teacher-led, and reflective structure uniquely positioned to generate actionable knowledge about instructional interventions in authentic classroom contexts (Han et al., 2023). However, a significant gap remains in the empirical literature: no published study to date has deliberately integrated AI-assisted features with a structured digital annotation framework as a combined intervention for developing critical reading of academic texts, nor has this combined intervention been investigated within a CAR methodological framework. Existing studies largely treat AI assistance and digital annotation as independent variables, leaving their synergistic potential uninvestigated. Furthermore, the majority of related studies employ quasi-experimental designs that, while internally rigorous, are less sensitive to the dynamic, contextual, and iterative dimensions of classroom learning that CAR is uniquely equipped to capture. The central theoretical tension that motivated this study was the conflict between AI's documented potential to scaffold higher-order analytical engagement and the empirically substantiated risk that unguided AI use produces cognitive offloading—a reduction of effortful cognitive engagement that may arrest rather than advance the development of critical reading competency. Addressing this tension requires not merely a new methodological combination but a theoretically principled pedagogical framework that deliberately structures AI use to promote, rather than replace, human analytical agency.

This study aimed to investigate the effectiveness of AI-assisted digital annotation as a Classroom Action Research intervention for improving undergraduate EFL students' critical reading of academic texts. Specifically, the study addresses three research questions: (1) How did AI-assisted digital annotation

improve students' critical reading of academic texts? (2) What changes in students' critical reading skills were observed across two cycles of classroom action research? and (3) How did students perceive the use of AI-assisted digital annotation in their academic reading process? The study was conducted with a class of 50 undergraduate students enrolled in Bahasa Inggris course at Universitas Dharma Indonesia, across two iterative CAR cycles guided by the Kemmis & McTaggart (1988) spiral model. Data were generated through Critical Reading Performance Tests and classroom observation, and were analysed using a mixed-method approach integrating descriptive and inferential statistical analysis with thematic analysis following Arini (2021). While the study's single-site, single-class design limits generalisability, its methodological rigour, contextual richness, and mixed-method triangulation enable meaningful transferability to comparable EFL higher education contexts (McCarty, 2020). The findings were intended to contribute empirically grounded insights to the growing scholarly discourse on AI integration in academic literacy instruction and to provide actionable pedagogical guidance for EFL practitioners seeking to improve students' critical engagement with academic texts.

## METHOD

### Research Type and Design

This study employed Classroom Action Research (CAR) as its methodological framework. CAR is a form of systematic, self-reflective, and practitioner-led inquiry conducted by teachers within their own classrooms with the dual purpose of solving immediate pedagogical problems and generating context-specific knowledge that can improve instructional practice with explicit epistemological justification for CAR, including acknowledgement of its comparative limitations relative to experimental designs and explicit restriction of causal claims to the study cohort (Vogelzang & Admiraal, 2017). It is particularly well-suited to the present study because the research problem—students' persistent difficulty in engaging critically with academic texts—was identified within the researcher's own instructional context, and the goal was to improve that specific practice through an iterative, evidence-informed cycle of intervention and reflection. The study adopted the spiral model proposed by Kemmis & McTaggart (1988), which organises the research process into four interconnected phases in each cycle: (1) Planning, (2) Acting, (3) Observing, and (4) Reflecting. Each cycle's reflection phase directly informs the planning of the subsequent cycle, producing a cumulative, iterative refinement of the instructional intervention. As Suhaimi & Musdizal (2022) noted, the spiral structure of CAR enables teachers to achieve progressive instructional improvement through successive cycles of reflective action—an advantage that quasi-experimental designs, constrained by pre-established fixed protocols, cannot replicate. The study was conducted across two cycles, each comprising four instructional sessions. Cycle I introduced and implemented the AI-assisted digital annotation intervention. The Cycle I reflection identified three specific instructional limitations, which were addressed through targeted refinements implemented in Cycle II. The decision to conduct two cycles was determined by whether the predetermined indicators of success—specifically, that at least 75% of students achieve a Critical Reading Performance Test score of ≥70—had been met at the conclusion of Cycle I; the shortfall in Cycle I (53.3% of students

achieving ≥70) necessitated the implementation of a revised Cycle II. A mixed-method research orientation was adopted, integrating quantitative pre-test and post-test data with qualitative data from observation, consistent with the methodological pluralism characterising contemporary CAR practice (Ismail & Fata, 2021).

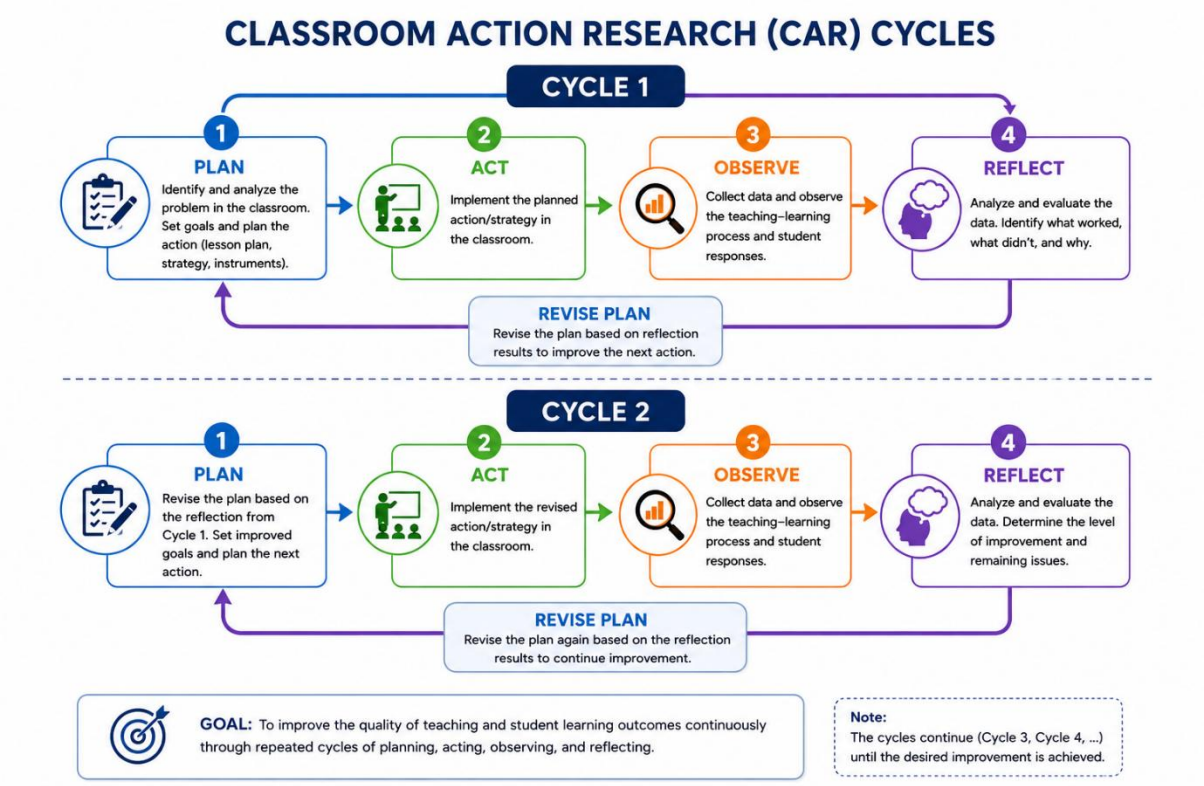



**Figure 1**. Classroom Action Research (CAR) Cycle Adapted from Kemmis & McTaggart

**Table 1.** Summary of Research Cycles

| Phase | Cycle I Activities | Focus | Cycle II Adjustments |
|---|---|---|---|
| Planning | Needs analysis; lesson plan design; preparation of annotation templates; orientation to AI tools | Establishing baseline; introducing AI-assisted annotation framework | Revised lesson plans based on Cycle I reflection; refined annotation prompts; strengthened AI scaffolding |
| Acting | Implementation of AI-assisted annotation across four sessions using two academic texts | Guided annotation with AI support; collaborative discussion | Implementation of revised intervention with additional metacognitive prompting; expanded peer annotation |
| Observing | Classroom observation; pre- | Monitoring student engagement and annotation quality | Systematic observation using refined checklist; |

| | | | |
|---|---|---|---|
| | /post-test administration | | collection of reflective journals |
| Reflecting | Analysis of Cycle I data; identification of gaps; planning for Cycle II | Evaluating whether success indicators were met | Final reflection on cumulative learning outcomes; assessment of overall intervention effectiveness |

**Sample and Research Subject**

The research was conducted in an undergraduate students of Universitas Dharma Indonesia. The participants comprised one class of [N = 50] undergraduate students enrolled in Bahasa Inggris course in the 2nd semester of 2025/2026 academic year. Participants were selected through purposive sampling, the standard sampling procedure in CAR research, whereby participants are naturally determined by the classroom context in which the teacher-researcher is working (Dewi et al., 2024). The class was selected for inclusion in the study on the basis of a preliminary diagnostic needs analysis, which identified a consistent pattern of difficulty in higher-order critical reading tasks—particularly argument evaluation, inference-making, and ideational synthesis—among the enrolled students. This finding was confirmed by the pre-test mean score of 49.8 (SD = 7.42), indicating that the overall class performance fell in the 'low' category prior to the intervention. All participants possessed intermediate-level English proficiency, as evidenced by their performance in the preceding English Language Skills course and their enrolment in the Bahasa Inggris course. Ethical procedures were observed throughout the study: written informed consent was obtained from all participants prior to data collection, anonymity and confidentiality of all participant data were maintained by replacing names with alphanumeric codes, and participants were informed of their right to withdraw from the study at any point without academic consequence.

**Instruments and Procedures**

Data were collected using two instruments, selected to enable methodological triangulation—the use of multiple data sources to cross-validate findings and enhance research credibility (Hafni & Fithriani, 2021). The instruments and their associated procedures are described below.

1. Critical Reading Performance Test (Pre-Test and Post-Test). A researcher-designed Critical Reading Performance Test was administered as a pre-test prior to the intervention and as a post-test at the conclusion of each cycle. The test measured performance across five critical reading dimensions: (a) identification of the author's main argument; (b) evaluation of supporting evidence; (c) inference-making from implicit information; (d) critical questioning and evaluation of assumptions; and (e) synthesis of ideas. Items were constructed as short-answer and extended-response questions requiring analytical engagement with previously unseen academic texts of comparable difficulty to those used in instruction. Content validity was established through expert review by two applied linguists specialising in EFL reading instruction, and inter-rater reliability for (a) explicit mapping of test items to Bloom's Revised Taxonomy and the Gebremariam & Weldeyohannes (2025) critical reading framework; (b) reporting of CVI (scale-level:

0.91); (c) reporting of inter-rater Kappa (0.83); (d) explicit acknowledgement of the rationale for a researcher-designed instrument.

2. Classroom Observation Checklist and Field Notes. Non-participant observation was conducted during all instructional sessions across both cycles using a structured observation checklist that recorded indicators of student engagement with annotation tasks, quality of AI tool use, and depth of critical analytical behaviour during class discussion. Narrative field notes supplemented the checklist to capture contextual details and micro-level classroom dynamics not accommodated by the structured instrument.

The instructional intervention itself—AI-assisted digital annotation—was implemented using a collaborative digital annotation platform (Hypothesis/Perusall) integrated with generative AI tools (ChatGPT) as scaffolding supports. Each cycle comprised four sessions of 90 minutes each, using selected peer-reviewed academic journal articles as reading texts. Students were guided through a five-component annotation framework requiring: (a) identification of main claims; (b) evaluation of evidence quality; (c) inference-making annotations; (d) questioning of assumptions with AI-assisted exploration; and (e) cross-textual synthesis. Cycle II incorporated three targeted refinements based on Cycle I reflection findings: enhanced metacognitive annotation prompts, explicit instruction in critical AI interrogation strategies, and structured collaborative synthesis tasks.

**Table 2.** Summary of Data Collection Instruments

| Instrument | Data Type | Purpose | Cycle |
|---|---|---|---|
| Critical Reading Performance Test | Quantitative | Measure CR performance gains | Pre + Post I + Post II |
| Observation Checklist & Field Notes | Qualitative | Document engagement & AI use quality | I & II |

**Data Analysis**

Data analysis followed a concurrent mixed-method design in which quantitative and qualitative analyses were conducted in parallel and then integrated at the interpretation stage to produce a comprehensive understanding of the intervention's effects and processes (Creswell, 2013).

1. Quantitative analysis was applied to the Critical Reading Performance Test data. Descriptive statistics—including means, standard deviations, and percentage gain scores—were calculated for overall scores and for each of the five critical reading dimensions at each test administration (pre-test, Cycle I post-test, Cycle II post-test). A paired-samples t-test was conducted to determine whether mean score improvements from pre-test to post-test within each cycle were statistically significant, with a significance threshold of $p < .05$. The practical significance of observed gains was assessed using Cohen's d effect size statistic, with values interpreted according to the conventional thresholds of small ($d \geq 0.2$), medium ($d \geq 0.5$), and large ($d \geq 0.8$) effects (Cohen, 1988). The percentage of students achieving the success threshold (≥70) was calculated at each post-test to assess whether the predetermined indicator of success had been met.

2. Qualitative analysis was applied to observation field notes, following the three-stage interactive model of Midgett et al. (2022): (1) data condensation—the systematic selection, paraphrasing, and summarisation of raw data to focus on information relevant to the research questions; (2) data display—the organisation of condensed data into thematic matrices, narrative summaries, and annotation quality profiles that made patterns and relationships visible; and (3) conclusion drawing and verification—the formulation of interpretive conclusions and their cross-verification against multiple data sources.

Trustworthiness was established through four strategies: (a) methodological triangulation—cross-validating findings across five instruments; (b) member checking—sharing interpretive summaries with participant sub-samples to verify accuracy; (c) audit trail documentation—maintaining records of all analytical decisions; and (d) researcher reflexivity—sustaining a reflective journal throughout the study to document and examine personal assumptions and emerging interpretive biases (Khatib et al., 2025).

## RESULTS AND DISCUSSION

This section presents and discusses the findings of the two-cycle CAR intervention, organised thematically around four crucial points: (1) the improvement of students' overall critical reading performance across cycles; (2) the development of dimension-specific critical reading sub-skills; (3) the emergence of metacognitive awareness and self-regulated reading strategies; and (4) students' perceptions of AI-assisted digital annotation. Each point presents empirical findings from the quantitative and qualitative data, followed by a discussion that situates those findings within the relevant scholarly literature. Quantitative data are drawn from the Critical Reading Performance Tests administered at three intervals (pre-test, Cycle I post-test, and Cycle II post-test), and qualitative data from classroom observation.

**Improvement in Overall Critical Reading Performance**

The quantitative results across both cycles demonstrated a pattern of significant and cumulative improvement in students' overall critical reading performance. Prior to the intervention, the class mean score on the Critical Reading Performance Test was 49.8 (SD = 7.42), with no student surpassing the 70-point threshold, indicating uniformly low baseline proficiency. Following the Cycle I intervention, the post-test mean rose to 64.7 (SD = 8.15), yielding a mean gain of 14.9 points. A paired-samples t-test confirmed that this improvement was statistically significant ($t(29) = 9.83$, $p < .001$), with a large effect size (Cohen's $d = 1.79$). However, only 53.3% of students ($n = 16$) achieved the predetermined success threshold of ≥70, falling short of the 75% target and necessitating Cycle II. Following the refined Cycle II intervention, the post-test mean increased further to 78.6 (SD = 7.23), a statistically significant improvement over the Cycle I post-test ($t(29) = 8.71$, $p < .001$, Cohen's $d = 1.59$) and a cumulative gain of 28.8 points from baseline. Critically, 83.3% of students ($n = 38$) achieved scores of ≥70 in the Cycle II post-test, surpassing the 75% success threshold and confirming that all predetermined indicators of success had been met. These gains were consistent across the student cohort, with improvements observed across the full performance spectrum from the lowest-performing to the highest-performing students.

**Table 3.** Comparison of Critical Reading Performance across Pre-Test, Cycle I, and Cycle II Post-Tests

| Measure | Pre-Test | Post-Test I | Post-Test II | Cumul. Gain |
|---|---|---|---|---|
| Mean Score (M) | 49.8 | 64.7 | 78.6 | +28.8 |
| Standard Deviation (SD) | 7.42 | 8.15 | 7.23 | — |
| Students Achieving ≥70 (%) | 0% | 53.3% | 83.3% | — |
| t-value (vs. previous) | — | 9.83*** | 8.71*** | — |
| **Cohen's d (effect size)** | — | 1.79 (large) | 1.59 (large) | — |

****p* < .001

The substantial and statistically significant gains in overall critical reading performance observed across both cycles are consistent with the findings of recent studies on strategy-based reading instruction in EFL contexts. Mekuria et al. (2024) demonstrated that structured reading strategy instruction over 12 weeks produced significant improvements in critical reading ability among Ethiopian EFL university students, a finding directly paralleled in the present study's two-cycle intervention. The superiority of Cycle II outcomes over Cycle I, with an additional 13.9-point mean gain following targeted pedagogical refinements, validates the fundamental premise of the CAR spiral model: that iterative cycles of reflective action produce cumulative instructional improvement beyond what a single-shot intervention can achieve (Bodner, 2016; Kemmis & McTaggart, 1988). The large effect sizes observed in both cycles (d = 1.79 and d = 1.59) are particularly noteworthy, suggesting that the AI-assisted digital annotation intervention produced improvements of meaningful practical significance well beyond the threshold of what is typically considered a large effect. These effect sizes are broadly consistent with the meta-analytic finding reported by Srinivasan & Murthy (2021) that digital annotation tools produce substantial positive impacts on reading comprehension when their use is guided by purposeful instructional frameworks, and with evidence from Hassan (2026) that AI-integrated instruction yields significant gains in EFL reading comprehension and academic engagement. The convergence of a structured annotation framework, AI scaffolding, and collaborative platform affordances within the present intervention may account for the magnitude of the observed effects, which exceed those typically reported for single-component interventions.

**Development of Dimension-Specific Critical Reading Sub-Skills**

Dimensional analysis of the performance test data reveals differentiated patterns of development across the five critical reading sub-skills measured, providing a more granular picture of the intervention's effects than overall mean scores alone. At baseline, the weakest dimensions were critical questioning and evaluation of assumptions (M = 44.6, SD = 9.1) and synthesis of ideas (M = 46.3, SD = 8.7)—both classified as 'very low'—while the relatively strongest dimension was identification of the main argument (M = 58.4, SD = 6.3). Following Cycle I, all five dimensions showed statistically significant improvements (all p < .001), yet the two weakest dimensions at baseline remained below the 70-point threshold (critical questioning: M = 58.4; synthesis of ideas: M = 60.2), indicating that the initial

intervention, while broadly effective, had not fully addressed the highest-order analytical demands. The Cycle II refinements—which specifically targeted metacognitive prompting and collaborative synthesis tasks—produced the most pronounced gains in precisely these two resistant dimensions: critical questioning improved by a further 18.0 points (Cycle I M = 58.4 → Cycle II M = 76.4) and synthesis of ideas by 15.6 points (Cycle I M = 60.2 → Cycle II M = 75.8). By the end of Cycle II, all five dimensions exceeded the 70-point mean threshold, with main argument identification achieving the highest mean (M = 82.3) and inference-making (M = 79.5) and evidence evaluation (M = 79.1) also performing strongly.

**Table 4.** Mean Scores by Critical Reading Dimension across All Test Administrations

| Critical Reading Dimension | Pre-Test M | Post-Test I M | Post-Test II M | Total Gain |
|---|---|---|---|---|
| Main Argument Identification | 58.4 | 72.1 | 82.3 | +23.9 |
| Evidence Evaluation | 51.2 | 63.1 | 79.1 | +27.9 |
| Inference-Making | 49.7 | 64.3 | 79.5 | +29.8 |
| Critical Questioning & Evaluation | 44.6 | 58.4 | 76.4 | +31.8 ↑ largest |
| Synthesis of Ideas | 46.3 | 60.2 | 75.8 | +29.5 |

The finding that higher-order critical reading dimensions—particularly critical questioning and synthesis—were the most resistant to change in Cycle I but showed the largest gains following targeted Cycle II refinements aligns closely with established accounts of the hierarchical nature of critical reading competency. Ravand et al. (2025) described critical reading as encompassing a hierarchy of sub-skills, with analysis and evaluation of text structures constituting cognitively more demanding operations than comprehension or identification tasks—a hierarchy directly reflected in the differential patterns of difficulty observed in the baseline and Cycle I data. The persistent difficulty with synthesis of ideas is particularly explicable through this framework: synthesis requires learners to move beyond engagement with a single text to construct relational understanding across textual sources, a task that demands the highest levels of analytical integration and is accordingly the most complex sub-skill within the critical reading taxonomy (van Ockenburg et al., 2019). The significant gains achieved in these dimensions following Cycle II's explicit metacognitive prompting and collaborative synthesis tasks suggest that the combination of structured annotation scaffolds and peer-mediated sense-making can successfully develop even the most demanding critical reading sub-skills, given adequate instructional design and sufficient iterative refinement. This finding supported the argument of Wu et al. (2021) that domain-specific reading strategy instruction, when structured around the cognitive demands of the target reading tasks, produces differentiated and meaningful improvements in the most challenging aspects of critical reading performance.

**Development of Metacognitive Awareness and Self-Regulated Reading**

Beyond quantitative performance gains, qualitative data from classroom observation, annotation artefacts, and student reflective journals documented a progressive and substantive development in students' metacognitive awareness

across both cycles. In Cycle I, early-session annotations were predominantly descriptive and paraphrastic in nature, reflecting surface-level processing of texts rather than analytical engagement. Observations showed that most students initially used AI tools in a passive query-and-paste manner, submitting text passages to AI assistants and accepting generated outputs without interrogation. The mean Annotation Quality Rubric score across all Cycle I artefacts was 2.1 out of 5.0 in Session 1, rising to 3.4 by Session 4, reflecting a gradual qualitative shift toward more evaluative and questioning annotation. Reflective journals from Cycle I documented students' growing awareness of their own comprehension limitations, with representative entries shifting from expressions of undifferentiated uncertainty ("I don't know if I understand this") to more metacognitively specific observations ("I understand the main argument but I'm not sure about the evidence for the third point"). In Cycle II, following the introduction of enhanced metacognitive annotation prompts, a qualitatively distinct self-monitoring pattern emerged: students were observed pausing during annotation tasks to revisit and revise earlier annotations, explicitly comparing AI-generated explanations against textual evidence before accepting them, and formulating their own analytical questions before consulting AI tools. By Cycle II Session 4, the mean Annotation Quality Rubric score had risen to 4.1, with the largest gains in the dimensions of 'metacognitive self-monitoring' (from 2.3 to 4.0) and 'critical interrogation of AI outputs' (from 2.8 to 4.2). Several students reported in Cycle II journals and interviews that they had developed a deliberate three-phase reading strategy: scanning the text to identify difficult or contestable passages; using AI to clarify and explore those passages; and then constructing independent analytical annotations based on their enriched understanding.

The metacognitive development documented in this study is consistent with a robust body of research linking metacognitive reading strategy instruction to improved comprehension and critical reading performance in EFL contexts. Yusuf et al. (2026) demonstrated that EFL learners who develop strong metacognitive strategies for planning, monitoring, and evaluating their reading processes—specifically, the three-phase model of pre-reading planning, mid-reading monitoring, and post-reading evaluation—show significantly superior comprehension and analytical performance compared to those who approach reading without metacognitive self-regulation. The three-phase reading strategy independently developed by students in Cycle II of the present study maps directly onto this framework, suggesting that the AI-assisted annotation intervention successfully scaffolded the internalisation of these metacognitive processes through structured practice rather than direct instruction. Fang & Rahman (2024) further established a significant positive correlation between EFL students' metacognitive awareness of reading strategies and their actual reading comprehension performance, providing theoretical grounding for the observed relationship between metacognitive growth and the quantitative test score gains documented in this study. The role of digital annotation in making metacognitive processes visible and practisable is also well-documented: Fu et al. (2026) confirmed that digital annotation tools facilitate metacognitive awareness development by requiring learners to externalise and articulate their comprehension monitoring through written commentary, thereby transforming tacit reading processes into explicit, observable, and improvable

cognitive acts. The convergence of structured annotation scaffolds, AI-assisted exploration, and explicit metacognitive prompting in Cycle II of this study appears to have been particularly effective in accelerating this developmental process.

**Students' Perceptions of AI-Assisted Digital Annotation**

Thematic analysis of semi-structured interview transcripts and reflective journals across both cycles produced four overarching themes describing students' perceptions of the AI-assisted digital annotation experience. First, the scaffolding function of AI was the most consistently reported perception: students characterised AI tools as reducing the anxiety and frustration previously associated with difficult academic texts by providing accessible vocabulary clarification, background contextualisation, and alternative explanations for complex arguments. A representative interview response from Cycle I illustrates this perception: "Before, when I couldn't understand a paragraph I just gave up. Now I ask [the AI] to explain it simply, and then I try to figure out what the author really means myself." Second, students reported substantial growth in metacognitive self-awareness, describing how the annotation framework had made visible and improvable cognitive processes previously experienced as opaque. Third, students expressed concerns about AI dependence and output reliability, noting instances where AI-generated explanations had been inaccurate or oversimplified, and where they had initially accepted these outputs uncritically. This concern was more prevalent in Cycle I data and markedly reduced in Cycle II following explicit instruction in AI interrogation strategies, with students describing habitual cross-checking behaviours by Cycle II: "Now I always check what the AI says against the text. If it doesn't match, I trust the text, not the AI." Fourth, students highlighted the social value of collaborative annotation, describing encounters with classmates' annotations as exposing them to divergent interpretations, motivating more careful analytical contributions, and creating a sense of intellectual community around the shared reading texts. Collaborative annotation interactions intensified markedly in Cycle II, with peer replies to annotations increasing in both frequency and analytical depth.

Students' perception of AI as a cognitive scaffold that reduces barriers to engagement while preserving space for independent analytical reasoning is consistent with the theoretical framework of AI as a Zone of Proximal Development mediator (Vygotsky, 1978). Fan (2025) documented similar scaffolding perceptions among EFL learners using AI-integrated reading platforms, finding that AI tools enhanced learner autonomy through flexible, real-time support while simultaneously cultivating students' confidence to engage with challenging texts. The concern about uncritical AI acceptance identified in Cycle I data mirrors a finding documented broadly in the AI-in-education literature: Hassan (2026) warned that without structured guidance, students' overreliance on AI may inhibit deep processing and reduce genuine cognitive investment in learning tasks. Crucially, the Cycle II finding that explicit instruction in AI interrogation substantially reduced uncritical acceptance behaviours is consistent with Feng et al. (2026) argumented that AI-integrated pedagogy must cultivate critical AI literacy as an explicit instructional goal, rather than assuming that critical AI engagement will develop spontaneously through exposure alone. Students' positive perceptions of collaborative annotation's social dimension are well-supported by the social constructivist literature: Xu & Dai (2025) found that social

annotation activities promoted strong cognitive and social engagement, with peer annotation interactions enriching individual analytical perspectives—a process that Thongkhotr & Chaijaroen (2026) conceptualized as a transformation of reading from a solitary to a collectively mediated sense-making activity, directly operationalising Vygotsky's (1978) principle that higher-order cognitive functions develop first in social interaction before being internalised as independent cognitive capacities.

## CONCLUSION

This Classroom Action Research study provides contextually grounded and methodologically triangulated evidence, within this specific classroom cohort that AI-assisted digital annotation is an effective intervention for improving EFL university students' critical reading of academic texts. Implemented across two reflective cycles guided by the Kemmis and McTaggart (1988) spiral model, the intervention produced a cumulative mean gain of 28.8 points from pre-test (M = 49.8) to the Cycle II post-test (M = 78.6), with 83.3% of students surpassing the predetermined success threshold of ≥70—exceeding the 75% target established at the outset of the study.

The intervention operated through three interlocking mechanisms: AI tools served as cognitive scaffolds that lowered barriers to engaging with complex academic texts; the structured annotation framework made critical reading sub-skills explicit and practised; and the collaborative annotation platform enriched analytical engagement through peer interaction. Qualitative findings corroborate these gains. Students reported marked growth in metacognitive awareness, adopting deliberate planning-monitoring-evaluation reading strategies consistent with the self-regulated reading behaviours documented by Martelletti et al. (2023). Equally significant was the emergence of critical AI literacy: by Cycle II, students routinely interrogated and cross-checked AI outputs against textual evidence—a disposition advocated by Ironsi & Bostanci (2026) as an essential 21st-century competency.

These findings carry clear pedagogical implications. AI tools must be embedded within structured, teacher-guided frameworks rather than used in an open-ended manner; and explicit instruction in critical AI interrogation is essential, not optional. While the single-site CAR design limits generalisability, the rich contextual documentation enables transferability to comparable EFL higher education settings. Future research should pursue longitudinal designs, controlled comparisons, and investigation of teacher professional development needs to broaden and deepen the evidence base. Ultimately, this study demonstrates that when AI is harnessed as a scaffold for critical thinking rather than a substitute for it, it can meaningfully improve students' engagement with the complex academic texts that undergraduate study demands.

## RECOMMENDATION

The recommendations presented in this section reflect a shared conviction, grounded in the empirical evidence of this study and the broader scholarly literature, that the integration of AI into academic reading instruction represents both a significant opportunity and a significant responsibility. The opportunity lies in AI's demonstrated potential to scaffold students' critical engagement with complex academic texts, cultivate metacognitive awareness, and foster collaborative reading

practices that extend and enrich individual analytical work. The responsibility lies in ensuring that this integration is designed, implemented, and governed in ways that develop rather than diminish students' capacity for critical, independent, and ethically grounded engagement with both academic texts and the AI tools that mediate their encounter with them. Meeting this responsibility requires coordinated action from teachers, curriculum designers, institutional leaders, and researchers working in concert toward a shared vision of AI-literate, critically engaged academic learners.

**Author Contributions**

The author was solely responsible for all aspects of this research, including conceptualisation, research design, data collection, classroom implementation of the AI-assisted digital annotation intervention, data analysis, interpretation of findings, and drafting and revising the manuscript in its entirety.

**Funding**

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The study was conducted as part of the author's independent academic inquiry within the institutional context of Universitas Dharma Indonesia.

**Acknowledgment**

The author thanks the student participants for their willing involvement in this study, the expert validators for their contributions to instrument development, and the anonymous reviewers for their valuable feedback on earlier drafts of this manuscript. Gratitude is also expressed to Universitas Dharma Indonesia for providing the institutional environment that made this research possible.

**Conflict of interests**

The author declares no conflict of interest. The research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## REFERENCES

Arini, Y. (2021). Improving Critical Reading Ability, Learning Autonomy, and Learning Participation Through Kahoot! Application. *LEKSEMA: Jurnal Bahasa Dan Sastra*, *6*(1), 91–100. https://doi.org/10.22515/LJBS.V6I1.3477

Aziz, M., & Rawian, R. (2022). Modeling higher order thinking skills and metacognitive awareness in English reading comprehension among university learners. *Frontiers in Education*, *7*, 1–17. https://doi.org/10.3389/feduc.2022.991015

Bodner, G. (2016). *Changing how data are collected can change what we learn from discipline-based educational research*. 3–10. https://doi.org/10.1119/perc.2016.plenary.001

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Lawrence

Erlbaum.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications.

Dewi, V., Degeng, P., & Razali, K. (2024). Exploring the students' perceptions of the integration of web-based and non-web-based ICT in ELT at the vocational high school in Indonesia. *Journal on English as a Foreign Language*, *14*(1), 285–315. https://doi.org/10.23971/jefl.v14i1.7818

Du, E. B., Nurkamto, J., & Suryani, N. (2024). Enhancing Critical Reading Through Metacognitive Scaffolding in Flipped-Classroom. *World Journal of English Language*, *14*(6), 297–307. https://doi.org/10.5430/wjel.v14n6p297

Fan, K. (2025). Enhancing Cognitive Reading Strategy Use for L2 Learners with Lower Working Memory through AI-Based Scaffolding. *Journal of English Language Teaching and Applied Linguistics*, *7*(3), 143–151. https://doi.org/10.32996/jeltal.2025.7.3.14

Fang, M., & Rahman, M. (2024). The Correlation between Metacognitive Strategies and IELTS Reading Achievement among Chinese Students. *Open Journal of Modern Linguistics*, *14*(3), 315–343. https://doi.org/10.4236/ojml.2024.143017

Feng, J. B., Diaz, G., & Long, Z. (2026). Learning in AI-Augmented Environments Through Dual Pathways of AI Literacy. *Research Square*, 1–24. https://doi.org/10.21203/rs.3.rs-9173238/v1

Fu, Y., & Hiniker, A. (2025). Supporting students' reading and cognition with AI. *The CHI '25 Workshop on Tools for Thought*. https://doi.org/10.48550/arXiv.2504.13900

Fu, Y., Wester, J., Berkel, N. Van, & Hiniker, A. (2026). Self-Regulated Reading with AI Support: An Eight-Week Study with Students. *ArXiv*, *1*(1), 1–26. https://doi.org/10.48550/arXiv.2602.09907

Gebremariam, H. T., & Weldeyohannes, M. Z. (2025). Cultivating students' reading skills: Inspiring reading comprehension and motivation through critical reading strategies. *SAGE Open*, *15*(2). https://doi.org/10.1177/21582440251367189

Hafni, A., & Fithriani, R. (2021). Integrating Wiki in Hybrid EFL Writing Class: Exploring Indonesian Students' Perceptions of Its Benefits and Challenges. *Journal of Community Research and Service*, *5*(1), 27. https://doi.org/10.24114/jcrs.v5i1.27186

Han, Z., Beccia, A., Creider, S., Lindhardsen, V., & Brinton, D. M. (2023). Teachers as Researchers : Exploratory Practice in Tunisian EFL Classrooms. *Studies in Applied Linguistics & TESOL*, *23*(2), 1–10. https://doi.org/10.52214/salt.v23i2.12025

Hassan, A. (2026). Leveraging artificial intelligence (AI) to enhance student engagement and academic performance in higher education. *Education and Information Technologies*. https://doi.org/10.1007/s10639-026-13946-w

Ironsi, C. S., & Bostanci, H. B. (2026). Enhancing the reading comprehension and autonomy of foreign language students using generative AI-integrated assistant. *Frontiers in Education*, *11*, 1–17. https://doi.org/10.3389/feduc.2026.1763238

Ismail, N. M., & Fata, I. (2021). Improving Reading TOEFL Score through Note Taking Strategy. *Al-Ta Lim Journal*, *28*(1), 46–54. https://doi.org/10.15548/jt.v28i1.668

Kemmis, S., & McTaggart, R. (1988). *The action research planner (3rd ed.)*. Deakin University Press.

Khatib, M., Shah, R., Mendhe, S., Whisner, C., & Buman, M. (2025). Bridging the divide: barriers and facilitators to equitable community-academic partnerships in health research. *Frontiers in Public Health*, *13*, 1617908. https://doi.org/10.3389/fpubh.2025.1617908

Kusumawardana, Z. R., & Akhiriyah, S. (2022). EFL university students' metacognitive awareness of reading strategy and its correlation with their reading comprehension. *Linguistics and ELT Journal*, *10*(1), 1–15. https://doi.org/10.33654/lej.v10i1.1543

Le, H. V., & Nguyen, L. Q. (2024). Promoting L2 learners' critical thinking skills: The role of social constructivism in reading class. *Frontiers in Education*, *9*, 1241973. https://doi.org/10.3389/feduc.2024.1241973

Li, M., Murphy, P. K., Wang, J., Mason, L. H., Firetto, C. M., Wei, L., & Chung, K. S. (2016). Promoting reading comprehension and critical–analytic thinking: A comparison of three approaches with fourth and fifth graders. *Contemporary Educational Psychology*, *46*, 101–115. https://doi.org/10.1016/j.cedpsych.2016.05.002

Martelletti, D. M., Luzuriaga, M., & Furman, M. (2023). 'What makes you say so?' Metacognition improves the sustained learning of inferential reading skills in English as a second language. *Trends in Neuroscience and Education*, *33*, 100213. https://doi.org/10.1016/j.tine.2023.100213

McCarty, R. (2020). *Towards a Theory and Practice of Translingual Transfer: A Study of 6 International Undergraduate Students*.

Mekuria, A., Bushisho, E. W., & Wubshet, H. (2024). The effects of reading strategy training on students ' reading strategy use and critical reading ability in EFL reading classes. *Cogent Education*, *11*(1), 2310444. https://doi.org/10.1080/2331186X.2024.2310444

Midgett, A., Doumas, D. M., & Buller, M. K. (2022). Posttraining Outcomes, Acceptability, and Technology-BasedDelivery of the STAC Bystander Bullying Intervention TeacherModule: Mixed Methods Study. *JMIR FORMATIVE RESEARCH*, *6*(8), 1–14. https://doi.org/10.2196/40022

Ravand, H., Shahi, R., Effatpanah, F., & Moghadamzadeh, A. (2025). Measuring cognitive levels in high-stakes testing: A CDM analysis of a university entrance examination using Bloom's Taxonomy. *Frontiers in Education*, *10*, 1–14.

https://doi.org/10.3389/feduc.2025.1644811

Srinivasan, V., & Murthy, H. (2021). Improving reading and comprehension in K-12: Evidence from a large-scale AI technology intervention in India. *Computers and Education: Artificial Intelligence*, *2*, 100019. https://doi.org/10.1016/j.caeai.2021.100019

Sublime, J., & Renna, I. (2024). Is ChatGPT Massively Used by Students Nowadays? A Survey on the Use of Large Language Models such as ChatGPT in Educational Settings. *ArXiv*, *arXiv:2412*, 1–40. https://doi.org/10.48550/arXiv.2412.17486

Suhaimi, S., & Musdizal, M. (2022). Implementation of the Consciousness-Raising (C-R) Method in Improving EFL Students' Writing Skills. *Tarbawi : Jurnal Ilmu Pendidikan*, *18*(1), 22–35. https://doi.org/10.32939/tarbawi.v18i1.1744

Tesfay, G., & Weldeyohannes, M. Z. (2024). The effects of reading strategy training on students' reading strategy use and critical reading ability in EFL reading classes. *Cogent Education*, *11*(1), 2310444. https://doi.org/10.1080/2331186X.2024.2310444

Thongkhotr, P., & Chaijaroen, S. (2026). Web-Based Constructivist Learning Environment Model to Foster Critical Reading and Reading Literacy within the PISA Framework : Comprehensive Development and Validation. *Journal of Education and Learning*, *15*(1), 428–445. https://doi.org/10.5539/jel.v15n1p428

van Ockenburg, L., Van Weijen, D., & Rijlaarsdam, G. (2019). Learning to write synthesis texts in secondary education: a review of intervention studies. *Journal of Writing Research*, *10*(3), 401–428. https://doi.org/10.17239/jowr-2019.10.03.01

Velásquez, E. (2025). Reading Between the Lines: Digital Annotation Insights from Heritage and L2 Learners. *Languages*, *10*(9), 207. https://doi.org/10.3390/languages10090207

Vogelzang, J., & Admiraal, W. (2017). Classroom action research on formative assessment in a context-based chemistry course. *Educational Action Research*, *25*(1), 155–166. https://doi.org/10.1080/09650792.2016.1177564

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Wu, L., Valcke, M., & Van Keer, H. (2021). Supporting struggling readers at secondary school: an intervention of reading strategy instruction. *Reading and Writing*, *34*(8), 2175–2201. https://doi.org/10.1007/s11145-021-10144-7

Xu, W., & Dai, W. (2025). "You are good annotators": Investigating how social reading based on social annotations and role assignment strategies facilitate learners' social interaction and knowledge construction. *Education and Information Technologies*, *30*(8), 10311–10341. https://doi.org/10.1007/s10639-024-13268-9

Yusuf, P., Pam, N., Nanpon, J., Yakubu, B., Naomi, N., & Samuel, G. I. (2026). The Impact of Generative AI Tutors on Metacognitive Monitoring and Self-Regulation in Higher Education : An Experimental Study. *The Journal of Contemporary*

*Academic Research and Methodologies (JCARM)*, *1*(2), 309–341. https://doi.org/10.5281/zenodo.19501204