From Assistance to Autonomy: AI Integration in Structured Research-Based Learning for Higher Education

Festiyed Festiyed[1], Desnita Desnita[2], Ziola Natasya[1], Muhammad Aizri Fadillah[1] and Fuja Novitra[2]

[1]Department of Science Education, Universitas Negeri Padang, Indonesia
[2]Department of Physics Education, Universitas Negeri Padang, Indonesia

festiyed@fmipa.unp.ac.id (corresponding author)
desnita@fmipa.unp.ac.id
ziolanatasya@student.unp.ac.id
m.aizrifadillah@gmail.com
fujanovitra@fmipa.unp.ac.id

https://doi.org/10.34190/ejel.24.1.4416

Abstract: Despite the growing interest in artificial intelligence (AI) for science education, little is known about its role within structured research-based learning (RBL) frameworks that balance technological assistance with developing independent research competencies. Existing studies often focus on AI as an isolated tool or a single-stage intervention, leaving a gap in understanding how AI can be systematically embedded across the research process without diminishing students' cognitive engagement. This study addresses that gap by implementing the newly developed IFTAR model, which organizes RBL into five sequential phases—Identification, Find Literature, Determine Methodology, Accommodate/Analyze/Interpret Data, and Report & Present—with AI selectively integrated into the literature search and data analysis stages. A quasi-experimental, non-equivalent control group PreTest–PostTest design was conducted with ninety undergraduate physics education students assigned to one control and two experimental groups. Cognitive outcomes were measured using a validated instrument and analyzed through classical ANCOVA, rank-based ANCOVA, and robust ANCOVA to account for assumption violations. Across all analytical approaches, both experimental groups significantly outperformed the control group, with no significant difference between the experimental conditions. These findings demonstrate that phase-specific AI integration within a transparent and scaffolded RBL framework can enhance cognitive performance while preserving methodological autonomy, offering a replicable model for purposeful AI use in STEM higher education.

Keywords: Artificial intelligence, Research-Based learning, IFTAR model, Physics education, Cognitive learning outcomes, Structured pedagogy

1. Introduction

Artificial intelligence (AI) has increasingly become a central component of higher education, transforming instructional practices and students' learning experiences across disciplines, including physics (Bitzenbauer, 2023; Festiyed et al., 2024). In the context of physics education, AI tools have been employed to support tasks such as literature searching, data analysis, simulation modeling, and the generation of instructional materials, functions that also align with the role of cognitive scaffolds that can enhance inquiry processes when appropriately guided (Fadillah, Usmeldi, & Asrizal, 2024; Linn et al., 2015; Sirisathitkul & Jaroonchokanan, 2025). Recent developments in generative AI, large language models (LLMs), have demonstrated capabilities in producing coherent and contextually relevant responses, which can aid in scaffolding students' inquiry and research-based learning (RBL) (Steinert et al., 2024; West, 2023). However, despite this promise, integration in formal physics curricula often remains fragmented, with AI applications limited to specific activities rather than embedded within coherent pedagogical designs (Kotsis & Vakarou, 2025; Leon, Lipuma, & Oviedo-Torres, 2025). This partial adoption may hinder the potential of AI to foster higher-order thinking, deep conceptual understanding, and sustained engagement in the learning process (Fadillah, Usmeldi, & Asrizal, 2024).

While AI adoption in education offers significant opportunities, it also presents challenges that require careful pedagogical consideration. Overreliance on AI-generated content can reduce students' opportunities for independent reasoning and critical thinking (Watts et al., 2023; Yik & Dood, 2024). Additionally, AI systems may provide linguistically fluent outputs but conceptually inaccurate outputs, posing risks to students' scientific understanding when used without appropriate guidance (Kortemeyer & Bauer, 2024; Sirnoorkar et al., 2024). Ethical issues such as plagiarism, data privacy, and bias in AI outputs further complicate integration in academic settings (Cotton, Cotton, & Shipway, 2024; Khowaja et al., 2024; Siregar et al., 2026). Furthermore, disparities in

Cite this article: Festiyed, F. et al. 2026. "From Assistance to Autonomy: AI Integration in Structured Research-Based Learning for Higher Education", *Electronic Journal of e-Learning*, 24(1), pp 109-124, https://doi.org/10.34190/ejel.24.1.4416

students' familiarity with AI tools and access to technological resources may lead to unequal learning experiences and exacerbate educational inequities (Agyare et al., 2025; Fadillah, Usmeldi, & Ravanis, 2025; Fadillah et al., 2026; Najdawi et al., 2024). These challenges emphasize the need for structured instructional approaches that balance AI's benefits with preserving students' active cognitive engagement. In light of these complexities, it becomes essential to anchor AI-supported instruction within robust theoretical foundations that can guide its pedagogical use. Research-based learning (RBL) positions students as active constructors of knowledge through iterative inquiry cycles (Brew, 2010; Healey, 2005), whereas scaffolding theory and cognitive apprenticeship emphasize modelling, guided participation, and the gradual release of responsibility to the learner (Collins, Brown, & Newman, 2018; Linn et al., 2015). From this theoretical perspective, AI can be conceptualized as a "conditional scaffold", a support mechanism that is activated selectively at cognitively demanding stages while ensuring that core reasoning and methodological autonomy remain with the student.

Building on these foundations, the present study investigates the integration of AI into a structured form of RBL within an undergraduate physics research methodology course. This instructional approach builds upon the pedagogical foundations of active, inquiry-oriented learning while introducing explicit scaffolding to guide students through the research process (Brew & Jewell, 2012; Brew & Saunders, 2020; Suyatman et al., 2021). Within this structure, the IFTAR model is used to organize the research process into sequential, transparent phases, clarifying where AI support is pedagogically appropriate and where independent reasoning must be preserved. By comparing cognitive outcomes between AI-supported structured research-based learning, the same model without AI, and traditional instruction, the study aims to provide empirical insights into how AI can be meaningfully embedded in physics higher education.

Accordingly, this study aims to examine how phase-specific AI integration within the IFTAR model influences students' cognitive learning outcomes, how these outcomes compare with those achieved through the same structured model without AI, and whether selective AI assistance can enhance learning without reducing students' autonomy in carrying out essential research tasks. This objective provides the basis for the following research questions:

RQ1: Does the integration of AI at selected stages of the IFTAR model improve students' cognitive learning outcomes compared to conventional instruction?

RQ2: How do the cognitive outcomes of students experiencing AI-supported structured RBL differ from those experiencing the same model without AI?

RQ3: To what extent does phase-specific AI integration support cognitive performance while preserving methodological autonomy?

2. Literature Review

2.1 Artificial Intelligence in Physics Learning

Research on AI in physics education has expanded rapidly, with recent work increasingly focused on how LLMs such as ChatGPT and GPT-4 shape students' conceptual understanding and problem-solving performance. Studies agree that LLMs produce fluent and structured explanations, yet they often contain subtle scientific inaccuracies or contradictory reasoning (Dahlkemper, Lahme, & Klein, 2023; Gregorcic & Pendrill, 2023). This tension—high linguistic quality but inconsistent epistemic reliability—recurs across the literature. For example, while Dao and Le (2023) and Tong et al. (2024) found GPT-4 performing at or above the level of many students on standard physics questions, Horchani (2025) and Kortemeyer (2023) showed that the same model struggles with context-rich problems requiring realistic assumptions and modelling. These contradictory findings suggest that AI performance is highly sensitive to task structure: LLMs excel when patterns are clear and representations familiar, but fail where domain reasoning must be constructed from physical principles.

Beyond accuracy, several studies examine how AI responses influence students' reasoning processes. Fadillah, Usmeldi, & Asrizal (2024) show that clarity, coherence, and conceptual links in AI outputs can support higher-order thinking in inquiry tasks. Yet these benefits materialize only when students engage critically with the output rather than accept it unexamined (Fadillah et al., 2025). This aligns with findings by Dahlkemper, Lahme, & Klein (2023), who report that students trust AI explanations even when errors are present, highlighting the importance of disciplinary literacy and evaluation skills.

A growing strand of research also raises concerns about bias, access, and academic integrity. Work by Bolukbasi et al. (2016) and Najdawi et al. (2024) underscores that AI systems can reproduce societal biases or exacerbate inequities when access to advanced tools is uneven. Similarly, Kortemeyer and Bauer (2024) warn that

unsupervised AI use may undermine authentic problem-solving efforts. Synthesizing these threads shows that the central challenge is not merely AI performance but how students and instructors manage AI as both a cognitive aid and a potential source of distortion. Thus, the literature points to the need for structured pedagogical frameworks that embed AI deliberately and ethically in physics learning.

2.2 Research-Based Learning and the Need for Structured Models

RBL is widely recognized for fostering deep engagement, authentic inquiry, and transferable research skills across STEM disciplines (Ward, Clarke, & Horton, 2014; Wessels et al., 2021). In physics, its alignment with disciplinary epistemic practices makes it particularly valuable, as students learn to design experiments, analyze data, and communicate results in ways consistent with professional scientific work (Docktor & Mestre, 2014; Ruf, Ahrenholtz, & Matthé, 2019). Numerous studies demonstrate that participation in RBL enhances students' analytical reasoning, methodological understanding, and confidence as emerging researchers (Bauer & Bennett, 2003; Lloyd, Shanks, & Lopatto, 2019; Russell, Hancock, & McCullough, 2007). However, this body of work also reveals important inconsistencies. While RBL is generally praised for promoting autonomy, several empirical studies show that insufficient structure can overwhelm students, especially in disciplines with high conceptual and methodological demands like physics (Estuhono, Festiyed, & Bentri, 2019; Redish, 2000; Thiem, Preetz, & Haberstroh, 2023). Wessels et al. (2021) highlight that the balance between independence and guided support is crucial: autonomy enhances ownership, but excessive independence before students achieve methodological readiness can lead to fragmented or superficial inquiry. This nuance is often overlooked in the literature, where autonomy is sometimes uncritically equated with authenticity.

Another layer of complexity emerges when considering educational levels. Thiem, Preetz, & Haberstroh (2023) demonstrate that undergraduate students benefit from exposure to the full research cycle for the first time, while master's students require deeper engagement with advanced analytical tools. Pourhejazy and Isaksen (2024) similarly argue that RBL must be adapted to students' disciplinary trajectories and prior experience. When taken together, these studies expose a fragmented landscape in which RBL is implemented inconsistently, often relying on instructors' interpretations rather than standardized or transparent structures. This highlights a growing need for RBL models that articulate steps clearly, support disciplinary progression, and maintain conceptual coherence.

2.3 Artificial Intelligence Within Structured Research-Based Learning Models

As AI becomes more present in higher education, researchers have begun examining how it can support different stages of RBL. AI tools can assist in literature searching, summarizing scientific texts, analyzing data, and visualizing results (Bitzenbauer, 2023; Bubeck et al., 2023; Hidayanto, Phusavat, & Kurnia, 2025; Woo, Guo, & Susanto, 2025). When used strategically, AI functions as a cognitive scaffold that frees students to focus on higher-order reasoning activities (Kortemeyer, 2023; Li, Huang, & Liu, 2024; West, 2023). However, current studies often examine isolated stages—such as literature review or data analysis—without considering how AI contributes across the full research cycle. As a result, the cumulative pedagogical impact of AI within RBL remains poorly understood.

Existing evidence also reveals contradictions. Some studies show that guided AI use enhances inquiry processes and improves students' interpretation of results (Dao et al., 2023; Hidayanto, Phusavat, & Kurnia, 2025). Others caution that unguided or excessive reliance on AI may distort reasoning or diminish essential cognitive effort (Gregorcic & Pendrill, 2023; Kortemeyer & Bauer, 2024). These contrasting findings suggest that the effectiveness of AI in RBL depends heavily on the structure in which it is embedded. Without clear boundaries and checkpoints, students may bypass core analytical processes or treat AI outputs as authoritative. Ethical considerations further complicate AI integration. Issues such as academic integrity, data privacy, transparency of AI-generated content, and algorithmic bias are increasingly emphasized in higher education research (Cotton, Cotton, & Shipway, 2024; Najdawi et al., 2024). While several studies recommend training and institutional guidelines, few explicitly connect these ethical tensions to students' development of epistemic responsibility—the ability to evaluate evidence, justify methodological decisions, and critique AI outputs. This gap reinforces the need for structured frameworks that integrate AI not only as a tool but as a catalyst for reflective and responsible inquiry.

2.4 Gaps in the Literature and Research Contribution

While integrating generative AI tools such as ChatGPT into science education has gained momentum, research examining their role within structured RBL frameworks remains limited. Studies such as Gregorcic and Pendrill (2023) have shown how ChatGPT can support conceptual reasoning in physics through Socratic-style dialogue,

yet also reveal its limitations in delivering nuanced explanations. Similarly, West (2023) and Bubeck et al. (2023) highlighted GPT-4's advanced reasoning capabilities, which could be leveraged to scaffold inquiry stages in RBL. Woo, Guo, & Susanto (2025) further demonstrated the potential of AI-driven feedback to enhance iterative refinement of student work, aligning well with the formative assessment dimension of RBL. These findings collectively suggest that AI could strengthen the inquiry process in science education; however, they also underscore the need for structured and purposeful integration within pedagogical models.

At the same time, existing research on RBL in higher education, particularly within science and engineering, has consistently demonstrated its benefits in fostering deep engagement, critical thinking, and authentic research skills (Bauer & Bennett, 2003;Russell, Hancock, & McCullough, 2007; Wessels et al., 2021). The foundational work of Brew (2010) and Healey (2005) emphasizes that RBL is most effective when students are engaged as active participants in the construction of knowledge, supported by a clear pedagogical structure. Nonetheless, the adaptation of RBL to online and AI-enhanced learning contexts remains underexplored. Emerging studies, such as those by Hosel et al. (2022), Hidayanto, Phusavat, & Kurnia (2025), Li, Huang, & Liu (2024), and Zawacki-Richter et al. (2019), indicate that AI can facilitate inquiry cycles, data analysis, and collaborative research in both physical and virtual environments. However, these studies generally focus on isolated stages of research rather than offering a comprehensive framework for AI integration throughout the process.

Existing RBL models, such as those outlined by Shaban, Abdulwahed, & Younes (2015), Tabuena et al. (2021), and Susiani, Salimi, & Hidayah (2018), typically present a sequence of research activities, including problem identification, literature review, methodology selection, data collection, analysis, and dissemination. While pedagogically valuable, these models often involve overlapping sub-stages, which can be difficult for novice students to follow. The challenge becomes even more pronounced in AI-integrated environments, where clearly defined boundaries between stages are essential for ensuring that AI tools are applied appropriately and do not inadvertently dominate or fragment the learning process. Without such clarity, students risk becoming overly reliant on AI without understanding the underlying cognitive and methodological steps in conducting research.

In response to these limitations, the present study introduces the IFTAR model, a streamlined adaptation of RBL designed to consolidate core research activities into five sequential phases: Identification of the problem and research question, Find Literature, Determine Methodology, Accommodate/Analyze/Interpret Data, and Report & Present. This reconfiguration draws from established RBL frameworks but intentionally reduces complexity, making the process more transparent and manageable for students. Importantly, it also provides clear entry points for integrating AI tools in ways that are aligned with constructivist and inquiry-oriented principles. By mapping specific AI functions to each IFTAR phase, the model seeks to balance technological support with the development of students' independent research competencies. A comparative visualization of traditional RBL frameworks and the proposed IFTAR model is provided to highlight these differences (see Figure 1).

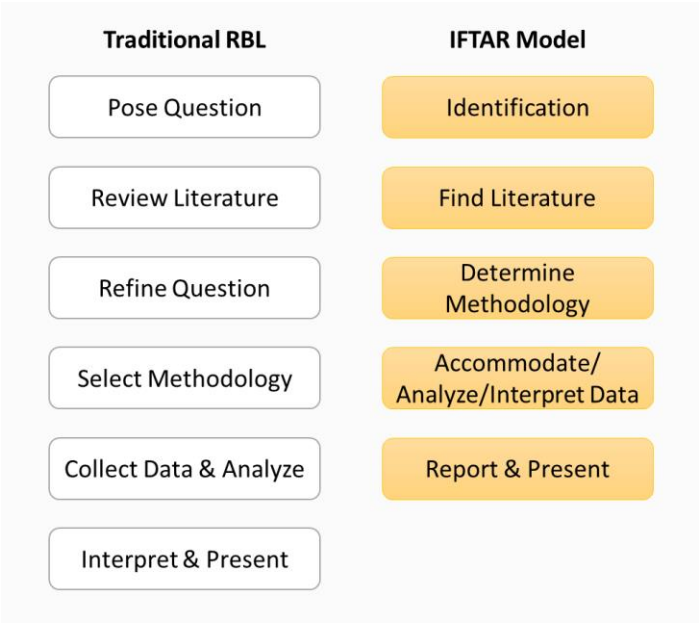

Figure 1: Comparison between traditional research-based learning and the IFTAR model proposed

Despite the potential advantages of such a structured approach, there remains a lack of empirical research that systematically examines how AI can be embedded in each phase of the IFTAR model. This gap is significant because, without phase-level integration strategies, the use of AI in RBL risks being fragmented, misaligned with intended learning outcomes, or ethically problematic. Furthermore, unresolved concerns regarding academic integrity (Cotton, Cotton, & Shipway, 2024; Kortemeyer & Bauer, 2024), algorithmic bias (Bolukbasi et al., 2016), and governance of AI in education (European Parliament, 2023; Najdawi et al., 2024) have yet to be addressed comprehensively in the context of student-led research. Therefore, this study aims to fill two critical gaps: mapping AI applications systematically across the IFTAR model stages and providing comparative insights into their pedagogical impact across educational levels and disciplinary contexts.

3. Methodology

This methodological design directly operationalizes the research gap and contribution outlined in Section 2.4, highlighting the lack of studies examining the pedagogical impact of AI integration at specific stages of structured research-based learning. The study was conducted in a higher education physics learning context, where integrating AI into structured, research-oriented learning can enhance students' critical thinking, autonomy, and conceptual understanding. The instructional model employed is the IFTAR model, representing the stages of Identification, Find Literature, Determine Methodology, Accommodate/Analyze/Interpret Data, and Report & Present. Adapted from the broader RBL framework (Shaban, Abdulwahed, & Younes, 2015; Susiani, Salimi, & Hidayah, 2018; Tabuena et al., 2021), IFTAR offers a more transparent and scaffolded structure, especially suited for embedding emerging technologies like AI. In this design, AI tools were strategically integrated into two specific stages, Find Literature and Accommodate/Analyze/Interpret Data, while the remaining stages were conducted without AI assistance to preserve students' independent reasoning and methodological autonomy. This partial integration approach allows a phase-specific analysis of AI's contribution to cognitive learning outcomes. The conceptual framework of the IFTAR model, including its AI integration points, is illustrated in Figure 2, showing clearly which stages are AI-assisted and which remain fully non-AI.

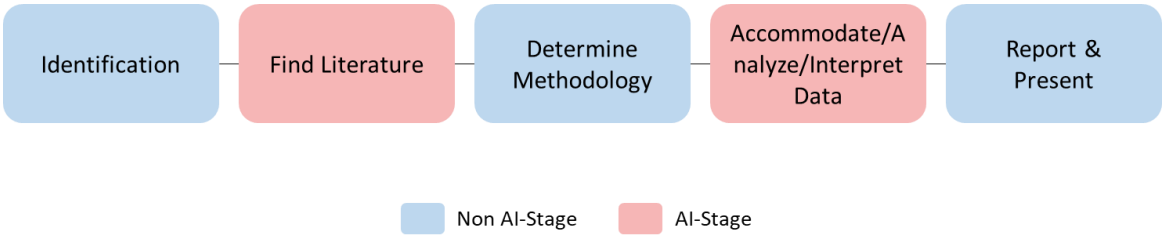

Figure 2: The IFTAR model stages highlight AI integration points. Blue sections represent stages conducted without AI assistance (Identification, Determine Methodology, and Report & Presenting), while red sections indicate AI-assisted stages (Find Literature and Accommodate/Analyze/Interpret Data)

3.1 Research Design

A quasi-experimental, non-equivalent control group PreTest–PostTest design was employed to evaluate the effectiveness of the IFTAR model with AI integration. Three intact classes participated in the study, each consisting of thirty undergraduate students, for a total of ninety participants. One class served as the control group and received conventional instruction, while the other two served as experimental groups taught using the IFTAR model with AI integration. All groups completed a PreTest at the beginning of the instructional period and a PostTest after learning. Figure 3 illustrates the study flow, from identifying participants through applying different instructional treatments and assessing outcomes via PostTest.

To reduce potential biases associated with non-equivalent intact classes, several control procedures were implemented. Baseline equivalence was examined by comparing PreTest scores and demographic characteristics, confirming no significant differences prior to the intervention. All groups were taught by the same instructor to eliminate variations in teaching style, and scheduling arrangements were structured to prevent treatment diffusion between groups. Random assignment at the individual level was not feasible due to institutional scheduling constraints, making intact class assignment the most ecologically valid approach for this quasi-experimental design.

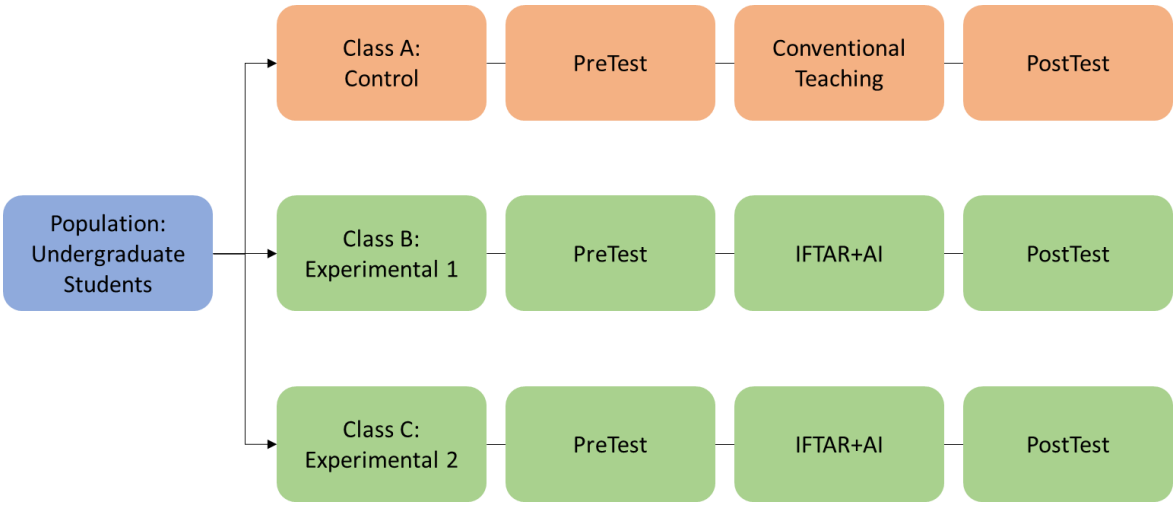

Figure 3: Research design flow from participant selection to instructional treatments and final assessment

3.2 Participants

The study involved ninety undergraduate students from the Physics Education program, all of whom were enrolled in the Research Methodology course during the semester of the study. The three intact classes, each consisting of thirty students, were assigned as follows: the control group (Class A) comprised 18 females and 12 males; Experimental Group 1 (Class B) included 19 females and 11 males; and Experimental Group 2 (Class C) consisted of 16 females and 14 males. This distribution resulted in fifty-three females and thirty-seven males across the sample. All participants had prior experience with AI tools, including platforms like ChatGPT, ensuring a uniform baseline familiarity with AI-supported learning across the sample. Additionally, students were well-accustomed to technology-enhanced learning environments, regularly engaging with digital platforms for materials, assignments, and collaboration. This consistency in technological background minimized variability in digital literacy, allowing differences in cognitive outcomes to be more confidently attributed to the instructional model implemented. The sample size of ninety participants also meets recommended guidelines for quasi-experimental designs with ANCOVA analysis. According to Cohen (2016), detecting a medium effect size (f = 0.25) at a power level of 0.80 with three groups requires a minimum sample of 66 participants. Similarly, Bujang and Baharum (2022) emphasize that a minimum of 20–30 participants per group for ANCOVA with one covariate is generally sufficient to achieve reliable statistical estimates. The allocation of 30 students per group in the present study ensures acceptable statistical power and is methodologically justified.

Ethical approval for the study was obtained from the Faculty Research Ethics Committee, and all participants provided informed consent prior to data collection. Participation was voluntary, and students were informed that their responses would remain confidential and would not affect course grades. Sampling followed a cluster-based approach, as students were already organized into intact course sections, which is appropriate for quasi-experimental designs in real instructional settings.

3.3 Intervention

The twelve-week intervention was implemented following the IFTAR learning model, distinguishing between AI-supported and traditional approaches. In the AI-supported approach, students can use various AI tools, such as ChatGPT, to support their process. All participants first completed a PreTest before Week 1. In Week 1, the control and experimental groups engaged in the Identification stage, defining research topics and formulating initial research questions. The control group followed instructor guidance and peer discussions, while the experimental group applied the same process but were informed of upcoming AI integration in later stages. Weeks 2–3 focused on the Find Literature stage. The control group performed manual searches using library databases, textbooks, and existing notes. The experimental groups, in contrast, used AI tools to generate keywords, locate relevant sources, and organize literature summaries, although final selection was made independently to maintain critical evaluation skills. In Weeks 4–5, all groups engaged in the Determine Methodology stage without AI assistance. They designed research procedures collaboratively, guided by instructor feedback, ensuring that methodological reasoning was entirely student-driven. Weeks 6–8 encompassed the Accommodate / Analyze / Interpret Data stage. The control group processed and interpreted data manually using a spreadsheet and statistical software, while the experimental groups utilised AI tools to

organise datasets, run preliminary analyses, and draft interpretations, retaining responsibility for verifying accuracy and validity. During Weeks 9–11, all groups participated in the Report & Present stage without AI support, preparing written reports and oral presentations to preserve originality and synthesis skills. Finally, in Week 12, all participants completed the PostTest to measure learning gains. Figure 4 illustrates this timeline, clearly marking the two stages where AI was integrated for experimental groups, while the control group followed traditional, non-AI-supported approaches throughout.

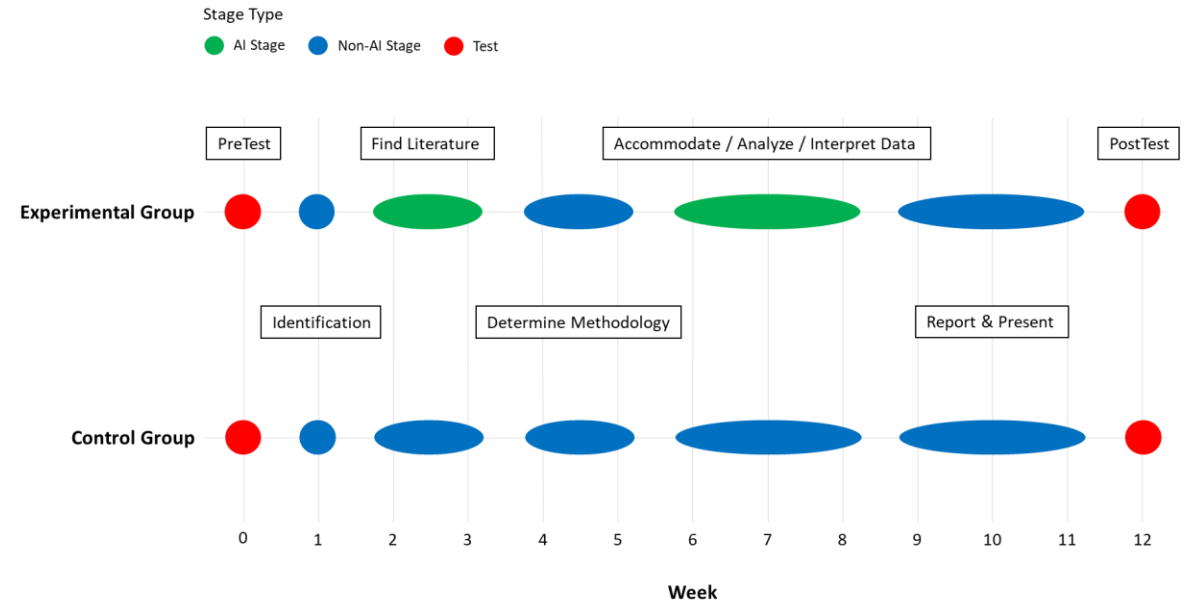

Figure 4: Timeline of intervention stages with AI integration

3.4 Cognitive Assessment Instrument

Cognitive learning outcomes were assessed using a test designed to align with the competencies targeted in each stage of the IFTAR learning model throughout the twelve-week intervention. The instrument was administered twice, once as a PreTest before the intervention and again as a PostTest upon its completion, to capture students' cognitive gains. It consisted of twenty items, comprising fifteen multiple-choice (MC) questions scored dichotomously (1 = correct, 0 = incorrect) and five open-ended (OE) tasks scored using a 0–5 analytic rubric reflecting accuracy, completeness, and reasoning quality.

The MC items measured factual and conceptual understanding of research methodology and educational inquiry. For example, one item asked: *"A student intends to investigate the effectiveness of simulation-based learning media. Which of the following should be the first step: developing a student satisfaction questionnaire, determining the number of respondents, constructing a theoretical framework from relevant literature, or clearly formulating the research problem?"* (correct answer: formulating the research problem). The OE tasks were designed to elicit higher-order thinking by placing students in authentic scenarios that required them to apply, analyze, and evaluate information. For instance: *"Explain the steps that should be taken to ensure that a research topic is relevant to the needs of physics education in schools?"*

The raw total score (maximum = 40) was transformed to a 0–100 scale to ensure comparability across formats. The instrument's content validity was established through an expert review by two specialists in physics education and research methodology, ensuring alignment with the intended learning outcomes for each stage of IFTAR. A pilot study involving fifteen respondents produced an internal consistency coefficient (Cronbach's alpha) of 0.745, which meets the generally accepted threshold of 0.70 for research purposes. Bujang, Omar, and Baharum (2018) show that a sample size of fewer than 30 can still provide reliable estimates for a single-coefficient alpha test when the minimum desired effect size is 0.70. This finding supports the adequacy of the pilot sample and confirms that the instrument demonstrates an acceptable level of reliability for measuring students' cognitive achievement.

3.5 Data Analysis

The data were analyzed using three complementary ANCOVA (analysis of covariance) approaches to ensure the robustness of the findings. The analysis began with assumption checks for the classical model, including Shapiro–

Wilk tests for residual normality and Levene's test for homogeneity of variance; the results indicated that these assumptions were not fully met (see Section 4.2 for more detail). First, a classical parametric ANCOVA was conducted in SPSS with PostTest as the dependent variable, PreTest as the covariate, and Group (Control, Experimental 1, Experimental 2) as the between-subjects factor, following standard procedures outlined by Field (2024). Second, a rank-based ANCOVA was performed in R to address the observed assumption violations by transforming the dependent variable into ranks before estimation, reducing sensitivity to distributional shape while retaining the interpretability of a linear model (Conover & Iman, 1981). Third, a robust ANCOVA in R using an M-estimator, implemented via the *lmrob* function in the *robustbase* package, was applied to down-weight influential observations and mitigate the effects of heteroscedasticity and outliers (Wilcox, 2011). These three analytical strategies allowed a thorough evaluation of the treatment effect's stability across methods with differing assumptions, ensuring that distributional irregularities or extreme cases did not unduly influence conclusions.

4. Results

4.1 Descriptive Statistics

Table 1 combines descriptive statistics for PreTest and PosTest so the reader can inspect baseline performance and outcomes after the intervention. For PreTest, group means were relatively similar: Control M = 52.00 (SE = 0.828, SD = 4.533, 95% CI [50.31, 53.69]), Experimental 1 M = 56.57 (SE = 2.527, SD = 13.843, 95% CI [51.40, 61.74]), and Experimental 2 M = 51.57 (SE = 2.561, SD = 14.026, 95% CI [46.33, 56.80]). These baseline values indicate no extreme initial advantage for any group, although Experimental 1 shows greater variance than the others. For PosTest, the pattern is more distinct: Control M = 60.60 (SE = 0.961, SD = 5.263, 95% CI [58.63, 62.57]), Experimental 1 M = 81.00 (SE = 2.377, SD = 13.017, 95% CI [76.14, 85.86]), and Experimental 2 M = 78.77 (SE = 2.573, SD = 14.093, 95% CI [73.50, 84.03]). The PosTest already points toward notable gains in the experimental groups relative to the control. Because raw PosTest differences do not account for baseline variability, ANCOVA (with PreTest as covariate) was applied to estimate adjusted group differences more accurately (see following subsections and Table 3–6).

Figure 5 presents a gain chart illustrating the mean PreTest and PostTest scores for each group to provide a more precise depiction of score progression. The lines connecting the points represent the average performance shift from the baseline to the post-intervention measurement. The control group showed a modest improvement, whereas both experimental groups displayed a substantial gain, with Experimental 1 exhibiting the largest increase. This pattern visually reinforces the descriptive statistics in Table 1, suggesting that the intervention methods implemented in the experimental groups were more effective than conventional instruction. Error bars in the figure represent 95% confidence intervals, allowing visual inspection of the precision of the mean estimates.

Table 1: Descriptive statistics for PreTest and PosTest by group

Group	N	PreTest Mean	PreTest SE	PreTest SD	PreTest 95% CI	PosTest Mean	PosTest SE	PosTest SD	PosTest 95% CI
Control	30	52.00	0.828	4.533	[50.31, 53.69]	60.60	0.961	5.263	[58.63, 62.57]
Experimental 1	30	56.57	2.527	13.843	[51.40, 61.74]	81.00	2.377	13.017	[76.14, 85.86]
Experimental 2	30	51.57	2.561	14.026	[46.33, 56.80]	78.77	2.573	14.093	[73.50, 84.03]

Note. The table shows means, standard errors (SE), standard deviations (SD), and 95% confidence interval (CI) as reported in SPSS output.

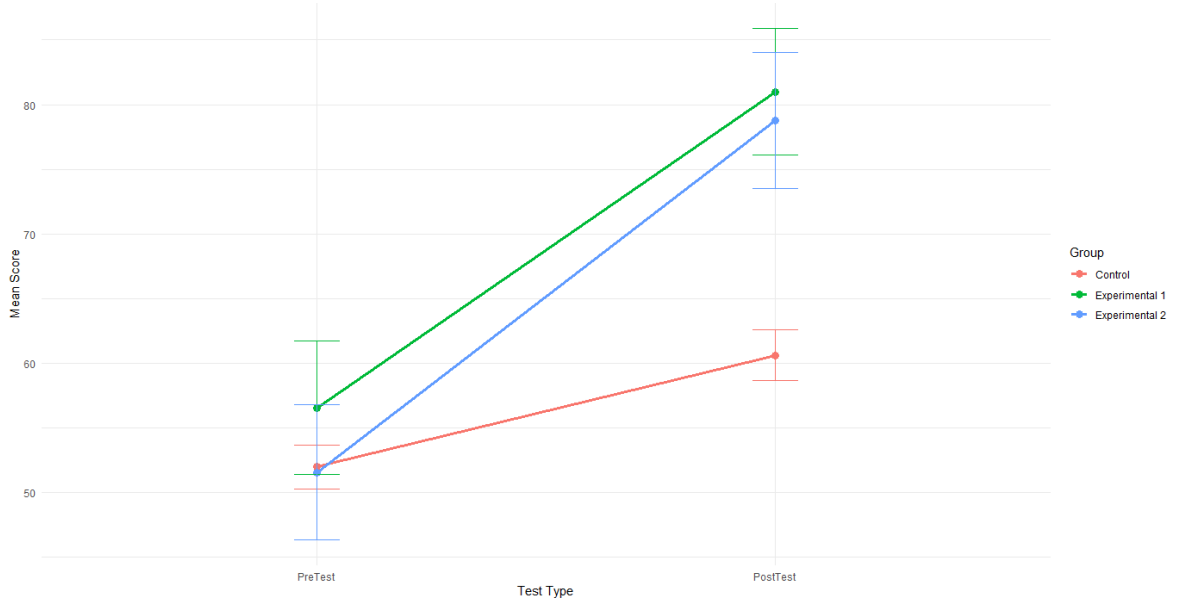

Figure 5: PreTest and PostTest mean scores by group

4.2 Assumption Checks

Before interpreting parametric ANCOVA, we evaluated key assumptions. Shapiro–Wilk tests for PosTest show deviations from normality in each group (Control W = 0.924, p = 0.034; Experimental 1 W = 0.909, p = 0.014; Experimental 2 W = 0.930, p = 0.048), indicating that residuals are not strictly normal at the group level. In contrast, PreTest showed no firm evidence of non-normality in SPSS (Control W = 0.971, p = 0.575; Experimental 1 W = 0.969, p = 0.509; Experimental 2 W = 0.939, p = 0.085), which supports its use as a covariate but does not remove concern for PosTest distributional shape. Levene's test for equality of error variances on PosTest was highly significant ($F(2,87) = 18.330$, $p < 0.001$), indicating heterogeneity of variances across groups. These results (non-normal PosTest distributions and heteroscedasticity) justify supplementing classical ANCOVA with rank-based and robust procedures to ensure violated assumptions do not drive inference. See Table 2 for the assumption-test summary.

Table 2: Assumption checks

Test	Group / Statistic	Value	p-value
Shapiro–Wilk (PosTest)	Control W	0.924	0.034
	Experimental 1 W	0.909	0.014
	Experimental 2 W	0.930	0.048
Levene's test (PosTest)	F(2,87)	18.330	< 0.001

4.3 Classical (Parametric) ANCOVA

The parametric ANCOVA (PostTest ~ PreTest + Group) is summarised in Table 3. The corrected model was significant: $F(3, 86) = 20.983$, $p < 0.001$, with $R^2 = 0.423$ (adjusted $R^2 = 0.402$). The covariate PreTest had a modest but statistically significant effect on post-test scores, SS = 513.202, $F(1,86) = 4.026$, $p = 0.048$, partial $\eta^2 = 0.045$, indicating baseline performance explained a small portion of variance in the outcome after controlling for Group. The main effect of Group was large and highly significant, SS = 7069.107, $F(2,86) = 27.726$, $p < 0.001$, partial $\eta^2 = 0.392$. Estimated marginal (adjusted) means at PreTest = 53.38 were: Control = 60.887 (SE = 2.066), Experimental 1 = 80.337 (SE = 2.088), and Experimental 2 = 79.143 (SE = 2.070). Bonferroni-corrected pairwise comparisons showed both experimental groups significantly exceeded the control (Experimental 1 − Control = 19.450, SE = 2.953, $p < 0.001$; Experimental 2 − Control = 18.257, SE = 2.916, $p < 0.001$), while the difference between Experimental 1 and Experimental 2 was not significant (mean diff = 1.193, SE = 2.961, $p = 1.000$). In short, the parametric ANCOVA—despite assumption concerns—shows a robust and large group effect favoring the two instructional interventions.

In addition, Figure 6 visualises the adjusted PostTest means derived from the ANCOVA model, which controls for differences in PreTest scores. The black dots represent the adjusted means, and the blue shaded bars indicate

the 95% confidence intervals. Red arrows between groups illustrate the direction of mean differences: an arrow pointing from one group to another indicates that the latter group has a higher adjusted mean score. In this study, arrows point from the Control group toward both Experimental 1 and Experimental 2, confirming that both experimental groups outperformed the control. Arrows are bidirectional between the two experimental groups, reflecting their nearly identical adjusted means and a nonsignificant statistical difference. This figure complements the numerical ANCOVA results, providing an intuitive visual confirmation of the intervention's positive effect.

Table 3: Classical ANCOVA

Source	SS	df	MS	F	p-value	partial η²
Corrected Model	8024.957	3	2674.986	20.983	< 0.001	0.423
PreTest	513.202	1	513.202	4.026	0.048	0.045
Group	7069.107	2	3534.554	27.726	< 0.001	0.392
Error	10963.365	86	127.481			
Adjusted Means (PreTest = 53.38)						
Group	**Mean**	**SE**				
Control (1)	60.887	2.066				
Experimental 1 (2)	80.337	2.088				
Experimental 2 (3)	79.143	2.070				
Pairwise comparisons (Bonferroni)						
Comparison	**Mean Difference**	**SE**	**p-value**			
Experimental 1 − Control	19.450	2.953	< 0.001			
Experimental 2 − Control	18.257	2.916	< 0.001			
Experimental 1 − Experimental 2	1.193	2.961	1.000			

Note. SS (Sum of Squares), df (degrees of freedom), MS (Mean Square), SE (standard error), F is the significance test statistic, and partial η² (partial eta squared) is the effect size that shows the proportion of explained variance.

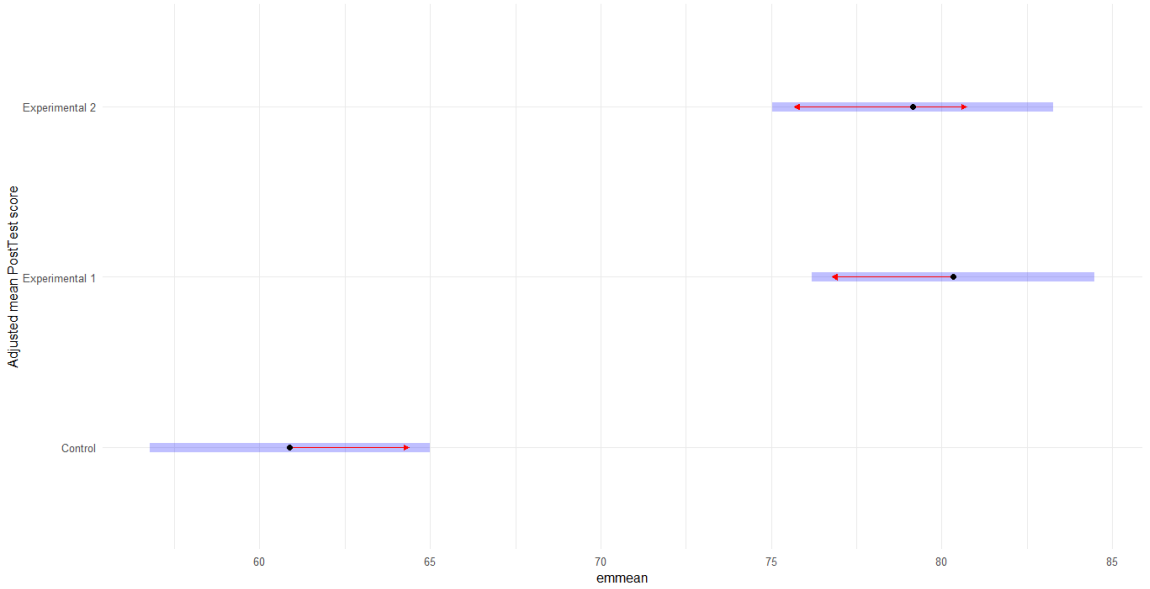

Figure 6: Adjusted PostTest means (ANCOVA)

4.4 Rank-Based ANCOVA

Because of the distributional violations observed above, we fitted a rank-based ANCOVA (rfit) in R, which operates on ranks and reduces sensitivity to non-normal shapes. The rank ANCOVA confirmed the presence of

a covariate effect: PreTest estimate = 0.2459 (SE = 0.1142), t = 2.1535, p = 0.034. Group contrasts were large and highly significant: Experimental 1 estimate = 20.1303 (SE = 3.2530), t = 6.1882, p < 0.001; Experimental 2 estimate = 19.0744 (SE = 3.2113), t = 5.9397, p < 0.001. The model's robust R^2 = 0.359 and the reduction-in-dispersion test (statistic = 16.03, p < 0.001) further indicate a consistent treatment effect that persists when analysis is performed on ranks. This result implies that the group differences are not driven solely by extreme values or skewness in the distributions—the ordering of scores across groups still favors the experimental conditions strongly. See Table 4 for the rank-ANCOVA coefficients.

Table 4: Rank ANCOVA

Predictor	Estimate	SE	t-value	p-value
(Intercept)	48.6664	6.4252	7.5743	< 0.001
PreTest	0.2459	0.1142	2.1535	0.034
Group: Experimental 1	20.1303	3.2530	6.1882	< 0.001
Group: Experimental 2	19.0744	3.2113	5.9397	< 0.001

4.5 Robust ANCOVA

An M-estimator-based robust ANCOVA (lmrob) was fitted to further protect inference from outliers and heteroscedasticity. The robust model again indicates strong group effects: Experimental 1 estimate = 19.4791 (SE = 3.0165), t = 6.457, p < 0.001; Experimental 2 estimate = 18.5714 (SE = 3.6466), t = 5.093, p < 0.001. Unlike the parametric and rank results, the covariate PreTest was not statistically significant in the robust fit (estimate = 0.25520, SE = 0.1824, t = 1.399, p = 0.165), suggesting that a few influential observations may have influenced the small covariate effect seen earlier. The robust model reports Multiple R^2 0.436 (adjusted = 0.417) and a robust residual standard error = of 10.34; robustness weights show that most observations received near-full weight while some were down-weighted (min weight = 0.533). The robust analysis supports the conclusion that the treatment effect is not an artifact of outliers and remains substantively important. See Table 5 for robust estimates.

Table 5: Robust ANCOVA

Predictor	Estimate	SE	t-value	p-value
(Intercept)	47.3843	9.5879	4.942	< 0.001
PreTest	0.25520	0.1824	1.399	0.165
Group: Experimental 1	19.4791	3.0165	6.457	< 0.001
Group: Experimental 2	18.5714	3.6466	5.093	< 0.001

4.6 Comparative Interpretation

Table 6 summarises the key conclusions across analytical approaches: parametric ANCOVA, rank-based ANCOVA, and robust ANCOVA. Across all three methods, the central finding is consistent: both Experimental 1 and Experimental 2 produced substantially higher post-test scores than the Control group, and the difference between Experimental 1 and Experimental 2 was negligible and not statistically significant. The main divergence concerns the role of the covariate PreTest: it was statistically significant in the classical parametric ANCOVA (p = 0.048) and in the rank ANCOVA (p = 0.034), but non-significant in the robust ANCOVA (p = 0.165). This pattern suggests that while baseline performance contributes to variance in some analytic frames, its effect is sensitive to a small number of influential observations or heteroscedastic errors; nonetheless, the group (treatment) effect is robust to these modeling choices. In substantive terms, we can be confident that the instructional interventions led to meaningful learning gains relative to conventional instruction, and that this conclusion holds under multiple estimation strategies that address different assumption concerns.

Table 6: Brief summary of analysis results

Analysis	Is PreTest significant?	Are groups significant?	Pattern
Parametric ANCOVA	Yes (p = 0.048)	Yes (p < 0.001)	E1 > Control; E2 > Control; E1 ≈ E2
Rank ANCOVA	Yes (p = 0.034)	Yes (p < 0.001)	E1 > Control; E2 > Control; E1 ≈ E2
Robust ANCOVA	No (p = 0.165)	Yes (p < 0.001)	E1 > Control; E2 > Control; E1 ≈ E2

5. Discussion

The present study set out to examine the pedagogical impact of integrating AI into a structured RBL framework, specifically the newly proposed IFTAR model, within an undergraduate physics research methodology course. The major finding is that both experimental groups, which adopted the IFTAR model with AI integration at two critical stages, literature searching and data analysis, achieved significantly higher post-test scores than the control group that received conventional instruction. This improvement was consistent across three complementary analytical approaches, classical ANCOVA, rank-based ANCOVA, and robust ANCOVA, thereby strengthening the robustness of the conclusions despite deviations from statistical assumptions. Notably, no statistically significant difference emerged between the two experimental groups, suggesting that the benefits of the AI integration were stable and did not depend on additional variations within the experimental treatment.

These findings underscore the importance of targeted AI integration within a well-defined pedagogical structure. Rather than allowing AI tools to permeate all aspects of the research process indiscriminately, the selective embedding of AI into specific phases of the IFTAR model appears to strike an effective balance between technological support and the preservation of students' independent cognitive engagement. It aligns with earlier research by Hidayanto, Phusavat, & Kurnia (2025) and West (2023), which suggested that AI is most beneficial when purposefully positioned within scaffolded learning activities rather than as a wholesale replacement for human reasoning. The current results extend these insights by providing empirical evidence at the whole-course level in physics education, demonstrating that even partial AI integration, when coupled with a transparent and sequential research framework, can yield substantial cognitive gains.

Beyond these empirical patterns, the findings also illuminate the underlying mechanisms through which AI supports learning within the IFTAR framework. Theoretically, the results are consistent with the principle of "augmented cognition," wherein technology amplifies, rather than replaces, human reasoning by reducing extraneous cognitive load during complex research tasks (Li, Huang, & Liu, 2024; Sweller, 2020). By embedding AI only at analytically intensive phases, the IFTAR model appears to operationalize a distributed cognition system (Hollan, Hutchins, & Kirsh, 2000), in which computational tools handle low-level processing while students retain responsibility for conceptual interpretation and methodological decisions. This alignment between technological affordances and cognitive task structure may explain why the experimental groups achieved higher-order gains without evidence of overreliance—a concern highlighted in prior studies (Gregorcic & Pendrill, 2023; Kortemeyer & Bauer, 2024). Conceptually, this positions the IFTAR model not merely as a procedural scaffold but as an advancement of existing RBL frameworks by providing a theory-based rationale for when and why AI should be integrated within inquiry cycles.

When viewed alongside previous studies, the outcomes of this research suggest both convergence and advancement. Similar to the work of Fadillah, Usmeldi, & Asrizal (2024) and Steinert et al. (2024), the integration of generative AI facilitated higher-order thinking skills, particularly in tasks requiring the synthesis of literature and the interpretation of complex datasets. However, whereas prior studies often examined AI as a single-stage intervention or as an auxiliary tool in isolated assignments, this study incorporated AI into a coherent pedagogical flow. Moreover, the structured nature of the IFTAR model mitigated common pitfalls identified in earlier evaluations, such as cognitive overload in open-ended inquiry tasks (Wessels et al., 2021) and overreliance on AI outputs without critical verification (Gregorcic & Pendrill, 2023). It provides a model for translating AI's potential into sustained learning gains, avoiding the fragmented adoption that has been criticized in the literature (Kotsis & Vakarou, 2025).

Although the results are encouraging, alternative explanations warrant consideration. The observed gains were partially influenced by novelty effects, wherein the introduction of AI tools generated heightened engagement independent of their cognitive utility. Students in the experimental groups might also have benefited from increased collaborative interactions during AI-assisted tasks, which could have contributed to learning gains irrespective of the AI's direct outputs. Furthermore, the instructor's role in guiding the AI-supported phases may have provided additional scaffolding that is not fully replicable in purely student-led contexts. While these factors do not diminish the observed group differences, they suggest that the benefits of AI integration may be intertwined with broader motivational and social dynamics in the classroom.

From a practical perspective, the findings hold several implications for educational design in physics and other STEM disciplines. The phase-specific AI integration demonstrated here offers a replicable model for institutions seeking to incorporate emerging technologies without undermining core disciplinary skills. In teacher education contexts, where graduates must both conduct and supervise research, structured AI-assisted RBL could prepare

future educators to leverage technology responsibly while fostering critical thinking among their students. Beyond physics, the model could be adapted for other domains that require literature synthesis, methodological rigor, and data interpretation, making it relevant for interdisciplinary and professional training programs.

Taken together, these findings highlight the study's contribution to advancing both the theory and practice of AI-supported inquiry learning. Conceptually, the IFTAR model provides a refinement of existing RBL frameworks (Brew & Jewell, 2012; Wessels et al., 2021) by articulating clearer phase boundaries and aligning them with specific cognitive functions where AI can serve as an epistemic partner rather than a procedural shortcut. Practically, the model offers a replicable structure for educators seeking to integrate AI responsibly without diminishing students' agency, addressing a gap frequently noted in the literature on fragmented or unguided AI adoption in higher education (Leon, Lipuma, & Oviedo-Torres, 2025; Zawacki-Richter et al., 2019). This dual contribution—conceptual clarification and practical design—demonstrates how AI-enhanced RBL can be systematically structured to maximize learning while preserving methodological autonomy.

Nevertheless, certain limitations must be acknowledged. The study was conducted within a single institution and involved a relatively homogenous sample of physics education undergraduates, which may limit the generalizability of the findings to other disciplines, educational levels, or cultural contexts. Although the sample size was adequate for the statistical analyses employed, larger-scale replications would allow for more nuanced subgroup analyses, such as gender differences or prior research experience. In addition, while cognitive learning outcomes were the primary focus, other dimensions such as long-term retention, metacognitive development, and attitudes toward AI-assisted learning were not measured, leaving important questions for future exploration.

Future research could build on these findings by examining the longitudinal effects of structured AI-supported RBL, particularly whether short-term cognitive gains evolve into sustained research competence and metacognitive growth over time (Linn et al., 2015; Thiem, Preetz, & Haberstroh, 2023). Such investigations would also clarify whether the phase-specific integration strategy proposed in the IFTAR model produces durable learning trajectories distinct from conventional AI-enhanced instruction. Comparative studies across disciplines and educational settings could test the adaptability of the IFTAR framework, while experimental variations could examine the optimal number and type of AI-assisted phases. Furthermore, qualitative investigations could provide richer insights into students' perceptions, strategies for validating AI outputs, and the socio-emotional aspects of engaging with AI in collaborative research environments. Such work would contribute to refining not only the pedagogical model but also the broader discourse on ethical, equitable, and effective AI integration in higher education.

6. Conclusion

This study demonstrates that artificial intelligence's strategic, phase-specific integration into a structured RBL framework, operationalized through the IFTAR model, can substantially improve students' cognitive outcomes while preserving essential research skills. By limiting AI assistance to the literature search and data analysis stages, the approach ensured that students benefited from technological efficiency without diminishing opportunities for independent reasoning and methodological decision-making. The consistent superiority of the experimental groups over the control group across multiple analytical approaches confirms that this balance between human agency and AI support is achievable and pedagogically effective. These findings highlight a clear direction for practice: for educators and curriculum designers, the key implication is that AI can be a powerful enabler in inquiry-driven learning when embedded purposefully within a transparent, scaffolded process. Unrestricted use risks replacing rather than enhancing student thinking, whereas deliberate, well-timed support can deepen understanding, foster autonomy, and prepare learners for the realities of research in AI-rich academic and professional environments. Future studies should explore how this model adapts to different disciplines, cultural contexts, and long-term learning trajectories, ensuring that AI integration functions as a catalyst for—not a substitute for—critical and creative inquiry.

Acknowledgements

The authors would like to express their deepest gratitude to the Ministry of Education, Culture, Research, and Technology (KEMDIKBUDRISTEK) of the Republic of Indonesia for funding this research as outlined in the master contract 088/C3/DT.05.00/PL/2025 and the derivative contract 2977/UN35.15/PL/2025, and the authors also sincerely thank all participants who contributed to this research.

AI Statement: Authors declare that artificial intelligence was not used to generate, analyze, or write the scientific content of this study. AI tools were used only for permissible tasks such as grammar checking.

Ethics Statement: This research involving human participants was approved by the Research Ethics Committee of Universitas Negeri Padang, which reviewed the methodology and confirmed adherence to ethical standards.

References

Agyare, B., Asare, J., Kraishan, A., Nkrumah, I., & Adjekum, D. K. (2025). A cross-national assessment of artificial intelligence (AI) Chatbot user perceptions in collegiate physics education. *Computers and Education: Artificial Intelligence*, *8*, 100365. https://doi.org/10.1016/j.caeai.2025.100365

Bauer, K. W., & Bennett, J. S. (2003). Alumni Perceptions Used to Assess Undergraduate Research Experience. *The Journal of Higher Education*, *74*(2), 210–230. https://doi.org/10.1080/00221546.2003.11777197

Bitzenbauer, P. (2023). ChatGPT in physics education: A pilot study on easy-to-implement activities. *Contemporary Educational Technology*, *15*(3), ep430. https://doi.org/10.30935/cedtech/13176

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, *29*.

Brew, A. (2010). Imperatives and challenges in integrating teaching and research. *Higher Education Research & Development*, *29*(2), 139–150. https://doi.org/10.1080/07294360903552451

Brew, A., & Jewell, E. (2012). Enhancing quality learning through experiences of research-based learning: implications for academic development. *International Journal for Academic Development*, *17*(1), 47–58. https://doi.org/10.1080/1360144X.2011.586461

Brew, A., & Saunders, C. (2020). Making sense of research-based learning in teacher education. *Teaching and Teacher Education*, *87*, 102935. https://doi.org/10.1016/j.tate.2019.102935

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., & Lundberg, S. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv Preprint ArXiv:2303.12712*. https://doi.org/https://doi.org/10.48550/arXiv.2303.12712

Bujang, M. A., & Baharum, N. (2022). Guidelines of the minimum sample size requirements for Kappa agreement test. *Epidemiology, Biostatistics, and Public Health*, *14*(2). https://doi.org/10.2427/12267

Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A Review on Sample Size Determination for Cronbach's Alpha Test: A Simple Guide for Researchers. *Malaysian Journal of Medical Sciences*, *25*(6), 85–99. https://doi.org/10.21315/mjms2018.25.6.9

Cohen, J. (2016). A power primer. In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (4th ed., pp. 279–284). American Psychological Association. https://doi.org/10.1037/14805-018

Collins, A., Brown, J. S., & Newman, S. E. (2018). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In *Knowing, learning, and instruction* (pp. 453–494). Routledge.

Conover, W. J., & Iman, R. L. (1981). Rank Transformations as a Bridge between Parametric and Nonparametric Statistics. *The American Statistician*, *35*(3), 124–129. https://doi.org/10.1080/00031305.1981.10479327

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, *61*(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

Dahlkemper, M. N., Lahme, S. Z., & Klein, P. (2023). How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT. *Physical Review Physics Education Research*, *19*(1), 010142. https://doi.org/10.1103/PhysRevPhysEducRes.19.010142

Dao, X.-Q., & Le, N.-B. (2023). Investigating the effectiveness of chatgpt in mathematical reasoning and problem solving: Evidence from the vietnamese national high school graduation examination. *ArXiv Preprint ArXiv:2306.06331*. https://doi.org/https://doi.org/10.48550/arXiv.2306.06331

Dao, X.-Q., Le, N.-B., Vo, T.-D., Phan, X.-D., Ngo, B.-B., Nguyen, V.-T., Nguyen, T.-M.-T., & Nguyen, H.-P. (2023). VNHSGE: VietNamese High School Graduation Examination Dataset for Large Language Models. *ArXiv Preprint ArXiv:2305.12199*. https://doi.org/https://doi.org/10.48550/arXiv.2305.12199

Docktor, J. L., & Mestre, J. P. (2014). Synthesis of discipline-based education research in physics. *Physical Review Special Topics - Physics Education Research*, *10*(2), 020119. https://doi.org/10.1103/PhysRevSTPER.10.020119

Estuhono, Festiyed, & Bentri, A. (2019). Preliminary research of developing a research-based learning model integrated by scientific approach on physics learning in senior high school. *Journal of Physics: Conference Series*, *1185*, 012041. https://doi.org/10.1088/1742-6596/1185/1/012041

European Parliament. (2023, June 8). *EU AI Act: First regulation on artificial intelligence*. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Fadillah, M. A., Tanjung, Y. I., Usmeldi, U., & Festiyed, F. (2025). Unveiling the role of critical thinking in education: regional analysis and predictive factors. *International Journal of Evaluation and Research in Education (IJERE)*, *14*(4), 2640–2651. https://doi.org/10.11591/ijere.v14i4.33234

Fadillah, M. A., Tarigan, M. R. M., Siregar, F. A., Amalia, L., & Usmeldi. (2026). How does technology support from teacher, parent, school, and learning affect students' learning motivation and overall performance Evidence from developing

countries. *International Journal of Mobile Learning and Organisation*, *20*(1), 93. https://doi.org/10.1504/IJMLO.2026.10071371

Fadillah, M. A., Usmeldi, U., & Asrizal, A. (2024). The role of ChatGPT and higher-order thinking skills as predictors of physics inquiry. *Journal of Baltic Science Education*, *23*(6), 1178–1192. https://doi.org/10.33225/jbse/24.23.1178

Fadillah, M. A., Usmeldi, U., & Ravanis, K. (2025). ICT-based physics learning: what activities are most important to predict students' confidence? *International Journal of Science Education*, 1–23. https://doi.org/10.1080/09500693.2025.2527377

Festiyed, F., Tanjung, Y. I., & Fadillah, M. A. (2024). ChatGPT in Science Education: A Visualization Analysis of Trends and Future Directions. *JOIV : International Journal on Informatics Visualization*, *8*(3–2), 1614–1624. https://doi.org/10.62527/joiv.8.3-2.2987

Field, A. (2024). *Discovering statistics using IBM SPSS statistics*. Sage publications limited.

Gregorcic, B., & Pendrill, A.-M. (2023). ChatGPT and the frustrated Socrates. *Physics Education*, *58*(3), 035021. https://doi.org/10.1088/1361-6552/acc299

Healey, M. (2005). Linking Research and Teaching to Benefit Student Learning. *Journal of Geography in Higher Education*, *29*(2), 183–201. https://doi.org/10.1080/03098260500130387

Hidayanto, A. N., Phusavat, K., & Kurnia, S. (2025). *Integrating Generative AI into Research-Based Learning for Undergraduate Students: Perceptions, Adoption Drivers, and Its Impact on Research Performance* (pp. 17–30). https://doi.org/10.1007/978-981-96-8430-4_2

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, *7*(2), 174–196. https://doi.org/10.1145/353485.353487

Horchani, R. (2025). ChatGPT's problem-solving abilities in context-rich and traditional physics problems. *Physics Education*, *60*(2), 025019. https://doi.org/10.1088/1361-6552/adb473

Hosel, C., Heinzig, M., Vogel, R., Roschke, C., Kuhn, A., Schmidsberger, F., Vodel, M., & Ritter, M. (2022). Adaptation of a Research-based Teaching-Learning Format with Approaches of Online Learning in the STEM Field. *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 1–6. https://doi.org/10.1109/ICECCME55909.2022.9988046

Khowaja, S. A., Khuwaja, P., Dev, K., Wang, W., & Nkenyereye, L. (2024). ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) Evaluation: A Review. *Cognitive Computation*, *16*(5), 2528–2550. https://doi.org/10.1007/s12559-024-10285-1

Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, *19*(1), 010132. https://doi.org/10.1103/PhysRevPhysEducRes.19.010132

Kortemeyer, G., & Bauer, W. (2024). Cheat sites and artificial intelligence usage in online introductory physics courses: What is the extent and what effect does it have on assessments? *Physical Review Physics Education Research*, *20*(1), 010145. https://doi.org/10.1103/PhysRevPhysEducRes.20.010145

Kotsis, K. T., & Vakarou, G. (2025). Bridging the Gap Between Newtonian and Einsteinian Thinking: The Impact of AI-Powered Teachers on Physics Education. *European Journal of Contemporary Education and E-Learning*, *3*(4), 32–45. https://doi.org/10.59324/ejceel.2025.3(4).03

Leon, C., Lipuma, J., & Oviedo-Torres, X. (2025). Artificial intelligence in STEM education: a transdisciplinary framework for engagement and innovation. *Frontiers in Education*, *10*. https://doi.org/10.3389/feduc.2025.1619888

Li, L., Huang, W., & Liu, B. (2024). A Research-Oriented Model for Artificial Intelligence Education: Integrating Multidisciplinary Approaches to Foster Innovation and Holistic Learning. *Proceedings of the 2024 7th International Conference on Educational Technology Management*, 480–485. https://doi.org/10.1145/3711403.3711481

Linn, M. C., Palmer, E., Baranger, A., Gerard, E., & Stone, E. (2015). Undergraduate research experiences: Impacts and opportunities. *Science*, *347*(6222). https://doi.org/10.1126/science.1261757

Lloyd, S. A., Shanks, R. A., & Lopatto, D. (2019). Perceived Student Benefits of an Undergraduate Physiological Psychology Laboratory Course. *Teaching of Psychology*, *46*(3), 215–222. https://doi.org/10.1177/0098628319853935

Najdawi, M. H. Al, Shwedeh, F., Abdelmoghies, M. M., Kitana, A., & Ali, A. (2024). Applying artificial intelligence in predicting educational excellence in higher education institutions: A case study in Jordanian universities. *Edelweiss Applied Science and Technology*, *8*(6), 7273–7289. https://doi.org/10.55214/25768484.v8i6.3579

Pourhejazy, P., & Isaksen, K. R. (2024). Exploring the curricular and pedagogical decision criteria for research-based learning design in undergraduate studies. *Evaluation and Program Planning*, *103*, 102409. https://doi.org/10.1016/j.evalprogplan.2024.102409

Redish, E. F. (2000). Discipline-Based Education and Education Research. *Journal of Applied Developmental Psychology*, *21*(1), 85–96. https://doi.org/10.1016/S0193-3973(99)00052-0

Ruf, A., Ahrenholtz, I., & Matthé, S. (2019). Inquiry-Based Learning in the Natural Sciences. In *Inquiry-Based Learning – Undergraduate Research* (pp. 191–204). Springer International Publishing. https://doi.org/10.1007/978-3-030-14223-0_18

Russell, S. H., Hancock, M. P., & McCullough, J. (2007). Benefits of Undergraduate Research Experiences. *Science*, *316*(5824), 548–549. https://doi.org/10.1126/science.1140384

Shaban, K., Abdulwahed, M., & Younes, A. (2015). Problem-centric process for research-based learning. *International Journal of Engineering Pedagogy (IJEP)*, *5*(2), 24–30.

Siregar, W. L., Usmeldi, U., Taali, T., & Fadillah, M. A. (2026). Tracing four decades of research on expert systems in engineering education: A bibliometric analysis. *Social Sciences & Humanities Open*, *13*, 102364. https://doi.org/10.1016/j.ssaho.2025.102364

Sirisathitkul, C., & Jaroonchokanan, N. (2025). Implementing ChatGPT as Tutor, Tutee, and Tool in Physics and Chemistry. *Substantia*, *9*(1), 89–101. https://doi.org/10.36253/Substantia-2808

Sirnoorkar, A., Zollman, D., Laverty, J. T., Magana, A. J., Rebello, N. S., & Bryan, L. A. (2024). Student and AI responses to physics problems examined through the lenses of sensemaking and mechanistic reasoning. *Computers and Education: Artificial Intelligence*, *7*, 100318. https://doi.org/10.1016/j.caeai.2024.100318

Steinert, S., Avila, K. E., Ruzika, S., Kuhn, J., & Küchemann, S. (2024). Harnessing large language models to develop research-based learning assistants for formative feedback. *Smart Learning Environments*, *11*(1), 62. https://doi.org/10.1186/s40561-024-00354-1

Susiani, T. S., Salimi, M., & Hidayah, R. (2018). Research Based Learning (RBL): How to Improve Critical Thinking Skills? *SHS Web of Conferences*, *42*, 00042. https://doi.org/10.1051/shsconf/20184200042

Suyatman, S., Saputro, S., Sunarno, W., & Sukarmin, S. (2021). The Implementation of Research-Based Learning Model in the Basic Science Concepts Course in Improving Analytical Thinking Skills. *European Journal of Educational Research*, *10*(3), 1051–1062. https://doi.org/10.12973/eu-jer.10.3.1051

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1–16. https://doi.org/10.1007/s11423-019-09701-3

Tabuena, A. C. T., Tabuena, Y. M. H., Lauber, Dr. O., & Chisag, Á. G. R. (2021). An Examination of the Effects of a Research-Based Instructional Model on Students' Critical Thinking Abilities in an Introductory Science Course. *International Journal of Research In Science & Engineering*, *11*, 1–12. https://doi.org/10.55529/ijrise.11.1.12

Thiem, J., Preetz, R., & Haberstroh, S. (2023). How research-based learning affects students' self-rated research competences: evidence from a longitudinal study across disciplines. *Studies in Higher Education*, *48*(7), 1039–1051. https://doi.org/10.1080/03075079.2023.2181326

Tong, D., Tao, Y., Zhang, K., Dong, X., Hu, Y., Pan, S., & Liu, Q. (2024). Investigating ChatGPT-4's performance in solving physics problems and its potential implications for education. *Asia Pacific Education Review*, *25*(5), 1379–1389. https://doi.org/10.1007/s12564-023-09913-6

Ward, J. R., Clarke, H. D., & Horton, J. L. (2014). Effects of a Research-Infused Botanical Curriculum on Undergraduates' Content Knowledge, STEM Competencies, and Attitudes toward Plant Sciences. *CBE—Life Sciences Education*, *13*(3), 387–396. https://doi.org/10.1187/cbe.13-12-0231

Watts, F. M., Dood, A. J., Shultz, G. V., & Rodriguez, J.-M. G. (2023). Comparing Student and Generative Artificial Intelligence Chatbot Responses to Organic Chemistry Writing-to-Learn Assignments. *Journal of Chemical Education*, *100*(10), 3806–3817. https://doi.org/10.1021/acs.jchemed.3c00664

Wessels, I., Rueß, J., Gess, C., Deicke, W., & Ziegler, M. (2021). Is research-based learning effective? Evidence from a pre–post analysis in the social sciences. *Studies in Higher Education*, *46*(12), 2595–2609. https://doi.org/10.1080/03075079.2020.1739014

West, C. G. (2023). Advances in apparent conceptual physics reasoning in GPT-4. *ArXiv Preprint ArXiv:2303.17012*. https://doi.org/https://doi.org/10.48550/arXiv.2303.17012

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Academic press.

Woo, D. J., Guo, K., & Susanto, H. (2025). EFL secondary students' use of ChatGPT for writing task completion pathways. *The Journal of Educational Research*, 1–14. https://doi.org/10.1080/00220671.2025.2510382

Yik, B. J., & Dood, A. J. (2024). ChatGPT Convincingly Explains Organic Chemistry Reaction Mechanisms Slightly Inaccurately with High Levels of Explanation Sophistication. *Journal of Chemical Education*, *101*(5), 1836–1846. https://doi.org/10.1021/acs.jchemed.4c00235

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0

