QUBAHAN ACADEMIC JOURNAL
VOL. 4, NO. 3, September 2024
https://doi.org/10.58429/qaj.v4n3a919

Evaluation of Problem Based Gamification Learning (PBGL) Model on Critical Thinking Ability with Artificial Intelligence Approach Integrated with ChatGPT API: An Experimental Study

Remerta Noni Naatonis [1], Rusijono [1], Miftakhul Jannah [1] and Edwin Ariesto Umbu Malahina [2] *

[1] Educational Technology Study Program, State University of Surabaya, Surabaya City, 60213, Indonesia;
[2] Informatics Engineering Study Program, STIKOM Uyelindo Kupang, Kupang City, 85111, Indonesia;

***Corresponding author*:** e-mail: edwinariesto@gmail.com.

ABSTRACT: This research evaluates the effectiveness of the PBGL model integrated with Artificial Intelligence (A.I) using the ChatGPT API to improve students' critical thinking skills. The main problems faced are the limited time of the lecturer in delivering the material, the lack of reference material, and the low motivation of students. This research used an experimental method involving 520 students from Timor Leste (Cova-Lima and Dili districts) and Indonesia (East Java and East Nusa Tenggara provinces) studying Python, Java, and Web programming. Participants were divided into control (n=260) and experimental (n=260) groups. Key challenges included integration of the ChatGPT API, implementation of A.I-based automated feedback, and maintaining consistency and fairness in testing. The evaluation results showed significant improvement in critical thinking skills in the experimental group across all topics. For the Python topic, the experimental group had an average improvement of 20.75 points, compared to 8.20 points in the control group, with an average difference of 12.55 points. On the Java topic, the experimental group increased by an average of 25.58 points, while the control group only 7.86 points, with an average difference of 17.72 points. In the Web Programming Languages topic, the experimental group showed an average increase of 23.74 points, compared to 8.58 points in the control group, with an average difference of 15.16 points. These findings confirm that the contribution of A.I and ChatGPT integration in the PBGL model can improve critical thinking, provide automatic feedback, and increase student motivation. This research has the potential to be a reference in digital curriculum design with generative A.I approaches such as ChatGPT API, showing that A.I can enrich learning materials, personalized materials and provide flexibility for students and teachers in the digital era.

Keywords: Gamification, Critical Thinking, ChatGPT API, Artificial Intelligence, PBGL.

I. INTRODUCTION

PBGL is an innovative approach to learning that combines game elements with problem-based learning methods. Reputable research shows that PBGL is effective in increasing learner motivation and engagement, as well as deepening understanding of complex concepts [1-3]. In PBGL, the learner is placed in a real or simulated situation that is challenging and requires the learner to apply learned knowledge and skills to solve a given problem in a variety of situations [4, 5]. Studies show that the use of game elements, such as points, levels, and challenges, can increase participants' active participation, encourage teamwork, and facilitate more meaningful and enjoyable learning [6-8]. Therefore, PBGL is one of the interesting and effective approaches in today's modern world of education, especially in overcoming the challenges of learning in the digital era and highly dynamic artificial intelligence.

The use of game elements in education, known as gamification, has grown rapidly since 2012 [9, 10]. Initially, gamification only involved simple things like quizzes and reward systems to make learning more engaging. However, this approach has evolved to become more complex, encompassing more sophisticated games to support deeper learning [11]. Today, gamification in education includes more complex scenarios and challenges that can adapt to real-life situations, thus helping learners to apply their knowledge in everyday life and encouraging educators to find new ways to make learning more interactive [12]. Alongside the development of gamification, the arrival of artificial intelligence in education has also opened up new opportunities for more personalized, intelligent, and automated learning experiences [13, 14]. A.I. technologies that can create and understand language, such as ChatGPT, have great potential to provide customized feedback to individual learners, as well as increase their engagement in learning [15]. A.I can customize lesson content to suit individual needs, respond to learner progress, and provide feedback that helps them think more critically [16]. This personalization ensures that learners stay engaged and get the tailored support they need to overcome barriers to learning and improve critical thinking skills [17, 18].

The purpose of this research is to see how effective the PBGL method is in helping learners think more critically. In this research, artificial intelligence technology such as ChatGPT API is used to enrich basic programming learning materials, such as Python, Java, and Web Programming. This technology provides freedom and customization for learners and teachers, so that learning can be better suited to the needs of each student. This research will also develop and test this PBGL model, and assess how learners can think critically by providing appropriate feedback based on learner progress. To evaluate the results, various statistical testing methods will be used, such as CVR, CV, homogeneity, normality, independent sample t-test, and gain score, which are basically ways to measure and analyze data obtained from students. Through this research, it is expected to find out how much impact the PBGL method has on students' critical thinking skills, and also what factors affect the success of applying this method in education.

The difference between the current research and previous research is more focused on the use of gamification to improve students' critical thinking skills in dealing with misinformation and fake news [19, 20]. Previous studies also discuss the use of learning platforms such as Kahoot! and Vonder Go. [21], as well as software with gamification elements such as points, levels and scores designed to motivate and engage learners [22, 23]. In addition, several studies have also highlighted the use of gamification in online discussions both individually and in communities [24, 25]. On the other hand, research on the use of artificial intelligence to support critical thinking more often discusses the anxiety of adopting A.I, as well as how A.I is used to measure and improve individual capabilities [26, 27]. There are also studies that assess academic integrity and learner attitudes towards using A.I. technologies [28], as well as looking at how A.I. affects critical thinking, motivation, and self-awareness [29]. Thus, this research will focus on evaluating how a learning model for basic programming such as Python, Java, and web programming integrated with ChatGPT's API-based intelligent system can help students and teachers. ChatGPT, as an A.I. engine, is able to generate texts that match the given material, thus providing flexibility and customizability in the learning process.

The research involved students from campuses in Timor Leste and Indonesia with a total of 520 participants, who were divided into two groups: a control group (n = 260) and an experimental group (n = 260). The participants were second and third semester students who had just learned basic programming languages such as Python, Java, and Web Programming Language. The courses tested were those basic programming languages, with a duration of three months that included 13 meetings. The average age of the students was 18 - 19 years old, and the testing involved pre-test and post-test assessments for each group. The selection of this course was based on previous evaluations that showed unsatisfactory scores in four Universities and Colleges, namely from Timor Leste (Cova-Lima and Dili districts) and Indonesia (East Java and East Nusa Tenggara provinces) according to interviews and surveys with the course instructors, with an average score of 65.5 and student distribution per class ranging from 10:20 to 7:23. The pass percentage in the previous semester only reached an average of 23.3%, which was caused by several factors, including time constraints, references to the material taught, as well as low enthusiasm, motivation, and desire to understand the material in depth. Therefore, this research aims to test students' critical thinking skills, enrich knowledge, and provide variety in basic programming materials by using an intelligent system-based gamification approach, as well as providing additional solutions and understanding for the students involved.

The problem in this research that needs to be considered is the integration of the system developed with the ChatGPT API in the context of PBGL for learning basic programming languages such as Python, Java and Web Programming Language which is still relatively new and requires thorough evaluation to ensure that this approach is truly effective in improving students' critical thinking skills. The technical challenges are related to the development and implementation of an A.I. model that can effectively provide formative feedback and an evaluation matrix based on student performance. In addition, this research involved two groups of students who were tested in different learning environments (control and experimental), so ensuring consistency and fairness in testing and evaluation was crucial. Lastly, the adaptability and flexibility of the A.I. model in meeting the individual needs of students and teachers must be ensured so that this approach can truly continue to be able to train critical thinking with system-generated questions with a given level of material and score, enriching additional knowledge and variety of material taught according to good academic standards. Overcoming these problems will be the key to the success of research in improving students' critical thinking skills through PBGL integrated with AI technology.

Based on the background and problem formulation that has been described, this research focuses on evaluating the effectiveness of PBGL with an A.I. technology approach integrated with ChatGPT API. The main objective of this research is to improve students' critical thinking skills through an innovative and adaptive learning model. PBGL that combines level and point elements with problem-based learning is expected to enrich the learning experience and deepen the understanding of complex concepts. Through an experiment involving the testing of basic programming materials such as Python, Java and Web Programming Language, this research will evaluate the impact of the approach on students' critical thinking skills, as well as identify key factors that influence the successful implementation of this model in an educational context. As such, the results of this research are expected to significantly contribute to the development of more effective and adaptive learning methods in the ever-evolving digital and artificial intelligence era.

II. LITERATURE REVIEW

This literature review aims to identify research gaps in studies regarding the integration of A.I and gamification in learning. While many studies have explored the benefits of A.I and gamification separately, there is limited understanding of their combined effectiveness in improving critical thinking skills. This review will analyze previous findings regarding the application of A.I and gamification in education, as well as explore under-explored areas that require further research to optimize A.I and gamification-based learning approaches.

1. THEORETICAL FRAMEWORK

Critical thinking is the ability to analyze, evaluate, and construct arguments based on logic and evidence. According to Kuhn (2019), critical thinking is a dialogical activity carried out by individuals, first interactively, then in a form that is internalized with others only implicitly [30]. In addition, the habit of critical thinking supports the development of analytical abilities necessary for solving complex problems and making informed decisions [31, 32]. Research by Butler (2024) found that critical thinking is a better predictor of wise life decisions than general intelligence [33]. This is reinforced by research by Zawacki-Richter et al. (2019) which showed that critical thinking skills allow participants to evaluate information objectively, crucial in today's digital age where false information can easily spread [34]. This is particularly important in an educational context, where the ability to think critically can help participants develop a deep understanding of the material or information being studied, improving understanding as well as learning outcomes.

PBL is a learning method in which students learn through solving real problems. In PBL, students are placed in challenging situations that require them to apply their learned knowledge and skills to solve the problem. This approach encourages active, collaborative and reflective learning. According to research by Chibueze and Theresa (2011), PBL increases students' intrinsic motivation and facilitates deeper mastery of concepts through direct engagement with the learning material [35]. In addition, PBL not only focuses on problem solving but also on developing critical and analytical thinking skills that are important in professional and academic life [36]. Laksmi et al. (2020) pointed out that PBL can address participants' various learning styles, making this approach more inclusive and effective in reaching a diverse population of participants [37]. PBL also allows students to work in groups, which strengthens their social and communication skills, as well as the ability to work together in teams to achieve a common goal [38, 39]. Sofie et al. (2017) added that PBL teaches students to learn independently and develop time management skills that are essential for success in higher education [40]. By placing students in real-world situations, PBL can help students see the relevance of what they are learning and how to apply it in a practical context, which ultimately increases engagement and motivation in the learning process.

Whereas, AI-driven feedback is a system that uses A.I. technology to provide automated and personalized feedback to participants [41]. This mechanism can analyze student performance in real-time and provide appropriate recommendations to improve collaboration and critical thinking [42]. The integration of ChatGPT with PBGL can enable more interactive and adaptive feedback, thus enriching students' learning experience [43].

2. *GAMIFICATION IN EDUCATION*

For decades, research on using game elements in learning (gamification) has been important in education for several reasons. In Asia, since 2020, gamification has focused on vocational schools, technical colleges, and healthcare [44]. Gamification is more suitable for use in problem solving [45] whose impact can improve the learning experience that was previously lacking and needs to be addressed again [46]. With the development of A.I, gamification approaches will be more adaptive [47]. The integration of A.I. and gamification in assessment has been shown to significantly improve participant engagement, motivation, and academic performance. This approach can make assessment more fun and impactful to learning, providing valuable feedback for educators to refine their pedagogical strategies [48] to mental health which holds great promise for improving one's emotional and psychological well-being [49].

The use of gamification in education has been shown to increase student motivation and engagement. Dichev and Dicheva (2017) showed that game elements such as points, badges and leaderboards can make the learning process more interesting and challenging for students [11]. This finding is in line with research by Alahmari (2020) who found that gamification in science education can significantly increase student engagement [50]. Then research by Smiderle et al. (2019) showed that gamification can help reduce learning distractions and improve student focus, but the integration of A.I. was not explored in depth. [51]. However, many of these studies focus on traditional gamification elements and lack consideration of the potential integration of more advanced A.I. technologies in an educational context.

3. *A.I IN EDUCATION*

The use of A.I in education has been explored to provide automated feedback and support adaptive learning. Chan and Zary (2019) showed that A.I. can impact personalized feedback and more efficient learning support [52]. Recent research by Bachiri and Mouncif (2020) showed that A.I systems can increase learner engagement in online courses through automated and adaptive feedback [53]. In addition, other studies highlighted that A.I., including ChatGPT, has great potential in providing personalized feedback in the context of distance education [54]. However, the use of APIs such as ChatGPT in the context of educational gamification is less explored. Existing research mostly discusses the use of A.I ChatGPT for general feedback without deeply integrating it into gamified learning systems [55, 56].

4. *GAMIFICATION AND A.I COMBINATION (RESEARCH-GAP)*

An existing research gap lies in the lack of exploration into the use of the ChatGPT API to provide insights and automated responses in gamification-based learning systems. Research on using A.I. to analyze learner behavior and customize gamification elements shows great potential [12]. Evaluation of test incorporation methods utilizing A.I techniques [57] examining the synergistic relationship between A.I. and gamification in relation to education, learning support systems, evaluation mechanisms, education management, and educators [58], as well as the impact of utilizing A.I in gamification [59] These are some of the areas that have been researched, but do not delve deeply into how APIs such as ChatGPT can be integrated to provide more personalized insights.

Thus, researchers Berg and Plessis (2023) recommended the need for further research to explore the practice of ChatGPT integration in lesson planning [60], as well as research by Lopez et.al which provides advice on the continuation of the utilization of ChatGPT in the critical thinking dimension [61]. This research was developed to understand how the integration of A.I and gamification can provide more personalized insights and influence long-term learning outcomes [62]. Existing studies are still limited to short-term outcomes and have not considered the long-term impact of A.I and gamification integration in education. Therefore, further research is needed to explore the potential of ChatGPT API in creating a more interactive, adaptive, and personalized learning environment, and to assess its effectiveness in improving students' engagement, motivation, and learning outcomes in the long term [63].

III. MATERIAL AND METHOD

In the research process carried out by researchers, it involves various methodological techniques and process flows used in the research which will be explained at the following points.

1. *RESEARCH DESIGN AND PARTICIPANT CRITERIA*

The research method used in this study is a quantitative method with a quasi-experimental design using a non-equivalent group control design technique. This research will conduct an experimental test, which is a way to find and observe the effect of something new or being tested directly, to get the truth in research. [64-66]. So, this research will be conducted directly to evaluate and analyze the testing of PBGL-based learning models involving students from campuses in Timor Leste and Indonesia with a total of 520 students from June 2023 - April 2024, which are divided into two groups: control group (n = 260) and experimental group (n = 260). The participants were second and third semester students who had just learned basic programming languages such as Python, Java, and Web Programming Language. The courses tested were those basic programming languages, with a duration of three months that included 13 meetings. The average age of the students was 18 - 19 years old, and the testing involved pre-test and post-test assessments for each group. The selection of this course was based on previous evaluations that showed unsatisfactory scores in four Universities and Colleges, namely from Timor Leste (Cova-Lima and Dili districts) and Indonesia (East Java and East Nusa Tenggara provinces) according to interviews and surveys with the course instructors, with an average score of 65.5 and student distribution per class ranging from 10:20 to 7:23. The pass percentage in the previous semester only reached an average of 23.3%, which was caused by several factors, including time constraints, references to the material taught, as well as low enthusiasm, motivation, and desire to understand the material in depth. Therefore, this research aims to test students' critical thinking skills, enrich knowledge, and provide variety in basic programming materials by using an intelligent system-based gamification approach, as well as providing additional solutions and understanding for students involved. The reason for choosing these criteria, where participants have previously understood basic programming languages such as Python, Java and Web Programming Language, age and education level suitability, balanced group distribution in 520 students, adequate research duration and pre-test and post-test testing in measuring the improvement of participants' understanding quantitatively.

The innovation in this research involves the development of a website service integrated with a generative ChatGPT API-based intelligent system. ChatGPT, as an AI engine, is able to generate text in natural language based on the context of the material or input provided by students, providing flexibility, adaptability, and ideas and insights to experimental group students. After students input answers or questions related to the material into the website system developed, the data is sent in JSON format to the ChatGPT API. ChatGPT API then processes this request based on the parameters requested by students, such as further explanation, answer correction, or programming code examples. The result produced by ChatGPT is text relevant to the request, such as in-depth explanations, correction suggestions, or additional insights. This output is then sent back to the website system, where students can view and use the information to improve their understanding or correct their answers. This system enables dynamic interaction between students and ChatGPT, providing a more adaptive and immersive learning experience.

2. DATA COLLECTION TOOLS

Data was collected using several tools or means, including pre-test and post-test to measure the initial and final knowledge and skills of 520 students in basic programming learning (Python, Java and Web Programming Language), scores and levels in the gamification system to evaluate students' performance during the gamification-based learning process, as well as automatic feedback/insight from A.I. using ChatGPT API to provide real-time feedback based on students' answers. The collected data was then analyzed using SPSS (Statistical Package for the Social Sciences) version 22 software, which allows in-depth statistical analysis and validation of research results. Using SPSS, this **research** was able to conduct various statistical tests such as CVR, CV, homogeneity, normality, independent sample t-test and gain score testing to ensure the accuracy and consistency of the data obtained.

3. ANALYSIS METHOD

The data analysis method was carried out using the inferential analysis method to test the hoptesis using the t-test, where the t-independent sample test was used to test for significant differences between the control group and the experimental group in the pre-test and post-test. Before testing the data hypothesis, it is necessary to test homogeneity using Levene to ensure the variance of the control and experimental group data is the same (homogeneous) with a p-value> 0.05. While normality testing uses Kolmogorov-Smirnova to show normally distributed data with a p-value> 0.05**.**

As for some of the tests used to test the validity and reliability of the data, namely; content validity test using lawshe technique (CVR) which evaluates the extent to which the items in the instrument cover all aspects measured by involving 8 experts [67]. Testing construction validity (CV) with factor analysis to ensure that the instrument can reflect the measured construct well. While reliability testing uses test-retest reliability to evaluate the consistency of the instrument in producing stable scores over time [68]. Data collection methods or techniques use critical thinking skills instruments with basic programming language materials such as Python, Java and Web Programming Language.

All statistical analyses were conducted using SPSS (Statistical Package for the Social Sciences) version 22 software, which allows validation of the research results through various in-depth statistical tests and ensures the accuracy and consistency of the data obtained.

4. GAMIFICATION-BASED PROGRAMMING LANGUAGE TOPIC

Gamification in other studies has had an impact that can improve academic performance, engagement, and motivation of learners [69] as well as having a psychological and intrinsic effect on participants [70, 71] gamification can be implemented in a variety of ways such as in focus group testing, online, utilizing software to developing concepts into the metaverse [72-75]. Gamification in this research is the application of game elements and principles in a non-game context to increase critical thinking, seriousness, insight and user engagement. Gamification is often used to make the learning process more interesting and interactive by adding elements such as points, levels, challenges and high scores. The aim is to increase the seriousness and motivation of learning and provide more fun and challenging feedback for students. The problem-based basic Python material that will be tested and entered into the system later to be tested can be seen in Table 1.

Table 1. Basic Python programming language gamification-based test materials with level and score/point elements

Level	Topic	Sub-Topic	Total Sub-Topic Question	Score per Question
Lv.1	Variables and Data Types	Variable declaration and initialization	10	10 - 100
		integer, float, string, boolean	10	10 - 100
		Data type conversion	10	10 - 100
		Basic operations on data types	10	10 - 100

Lv.2	Operators and Expressions	Arithmetic operators	10	10 - 100
		Assignment operator	10	10 - 100
		Comparison operator	10	10 - 100
		Logical operators	10	10 - 100
Lv.3	Branching and Looping	if, elif, and else	10	10 - 100
		Looping with for	10	10 - 100
		Looping with while	10	10 - 100
		break, continue, and pass	10	10 - 100
Lv.4	Data Structure	List	10	10 - 100
		Tuple	10	10 - 100
		Set	10	10 - 100
		Dictionary	10	10 - 100

The data in Table 1 describes the structure of the materials and sub-materials that will be tested on students in learning basic Python programming language, organized by level. Each level has one main topic, and this topic is broken down into 4 sub-matters. Each sub-matter consists of 10 questions that students must answer. For each correct answer, students will get a score of 10, while incorrect answers will not get a score (score 0). The testing process starts with the first sub-matter in level 1. Students must answer all 10 questions in this sub-matter first. The number of questions in each sub-matter can be adjusted by the instructor as needed. After the first sub-matter is completed, students will proceed to the second, third, and fourth sub-matter. After all sub-matter in level 1 is completed, students will move on to the next level, which is level 2, and so on until level 4. Once the student has completed all the questions in all the sub-matter at each level, the final score will be displayed on the system. This allows students to see real-time results and progress after completing the gamification of the Python material.

Furthermore, the material about Java, which is an object-oriented programming language, where Java has a clear syntax and is relatively easy to understand by the test students, where Java is also taught at 4 universities and colleges that are the object of research. The structure of Java material is divided into 4 levels and also each level has 4 sub-topics of tested material. The basic problem-based Java material that will be tested and entered into the system later to be tested can be seen in Table 2.

Table 2. Basic Java programming language gamification-based test materials with level and score/point elements

Level	Topic	Sub-Topic	Total Sub-Topic Question	Score per Question
Lv.1	Variables and Data Types	Variable declaration and initialization	10	10 - 100
		integer, float, string, boolean	10	10 - 100
		Type casting and conversion	10	10 - 100
		Basic operations on data types	10	10 - 100
Lv.2	Operators and Expressions	Arithmetic operators	10	10 - 100
		Assignment operator	10	10 - 100
		Comparison operator	10	10 - 100
		Logical operators	10	10 - 100
Lv.3	Control Flow Statements	if, elif, and else statements	10	10 - 100
		Looping with for	10	10 - 100
		Looping with while	10	10 - 100
		Break, continue, and return statements	10	10 - 100
Lv.4	Object-Oriented Programming	Classes and Objects	10	10 - 100
		Inheritance	10	10 - 100

		Polymorphism	10	10 - 100
		Encapsulation and Access Modifiers	10	10 - 100

Table 2 describes the structure of the materials (Variables and Data Types, Operators and Expressions, Control Flow Statements and Object-Oriented Programming) and sub-materials to be tested on students in learning basic Java programming language, organized by level. Each level has one main topic, and this topic is broken down into 4 sub-matters. Each sub-matter consists of 10 questions that students must answer. For each correct answer, students will get a score of 10, while incorrect answers will not get a score (score 0). The testing process starts with the first sub-matter in level 1. Students must answer all 10 questions in this sub-matter first. The number of questions in each sub-matter can be adjusted by the instructor as needed. After the first sub-matter is completed, students will proceed to the second, third, and fourth sub-matter. After all the sub-matter in level 1 is completed, students will move on to the next level, which is level 2, and so on until level 4. Once the student has completed all the questions in all the sub-matter at each level, the final score will be displayed on the system. This allows students to see real-time results and progress after completing the gamification of the Java material.

The last material is Web Programming Language, which has several advantages that make this programming language very popular in the development of applications and websites. This language is easy to learn and use, making it suitable for beginners. Web Programming Language is also a course that is certainly required to be taught on every Informatics-based campus in the world, and is also a compulsory course taught at the 4 universities and colleges that are the object of research. The basic Web Programming Language material structure is divided into 4 levels and also each level has 4 sub-topics of material tested. The problem-based basic Web Programming Language material that will be tested and entered into the system later to be tested can be seen in Table 3.

Table 3. Basic Java programming language gamification-based test materials with level and score/point elements

Level	Topic	Sub-Topic	Total Sub-Topic Question	Score per Question
Lv.1	HTML Basics	Structure of HTML Document	10	10 - 100
		Common HTML Tags (headings, paragraphs, links)	10	10 - 100
		Images and Multimedia Embedding	10	10 - 100
		Lists (ordered, unordered)	10	10 - 100
Lv.2	CSS Basics	CSS Syntax and Selectors	10	10 - 100
		Box Model and Layout	10	10 - 100
		Comparison operator	10	10 - 100
		Responsive Design with Media Queries	10	10 - 100
Lv.3	JavaScript Fundamentals	JavaScript Syntax and Variables	10	10 - 100
		Functions and Events	10	10 - 100
		DOM Manipulation	10	10 - 100
		Basic Form Validation	10	10 - 100
Lv.4	Advanced JavaScript	Arrays and Objects	10	10 - 100
		Loops and Iteration	10	10 - 100
		Error Handling (try-catch)	10	10 - 100
		Introduction to Asynchronous JavaScript (Promises, async/await)	10	10 - 100

Table 3 describes the structure of the materials (HTML Basics, CSS Basics, JavaScript Fundamentals and Advanced JavaScript) and sub-materials that will be tested on students in learning basic Web Programming Language, organized by level. Each level has one main topic, and this topic is broken down into 4 sub-matters. Each sub-matter consists of 10 questions that students must answer. For each correct answer, students will get a score of 10, while incorrect answers will not get a score (score 0). The testing process starts with the first sub-matter in level 1. Students must answer all 10 questions in this sub-matter first. The number of questions in each sub-matter can be adjusted by the instructor as needed. After the first sub-matter is completed, students will proceed to the second, third, and fourth sub-matter. After all the sub-matter in level 1 is completed, students will move on to the next level, which is level 2, and so on until level 4. Once the student has completed all the questions in all the sub-matter at each level, the final score will be displayed on the system. This allows students to see real-time results and progress after completing the gamification of the web material.

5. CHATGPT INTEGRATED A.I-BASED GAMIFICATION MODEL ARCHITECTURE

The framework or gamification system model to be implemented is designed to fulfill some of the main elements of a game, such as basic programming materials (Python, Java, and Web Programming Language), levels, and scores. This model will be developed and implemented based on artificial intelligence (AI). In this system, each correct answer will be automatically detected by AI, which then provides additional insights and explanatory material to students. The AI also enables the measurement of student satisfaction through ratings of the materials and insights provided, using a score scale of 1 to 5. These ratings provide instructors with important feedback for further evaluation of the system's effectiveness. The framework of this AI system integrated with the ChatGPT API can be seen in Figure 1.

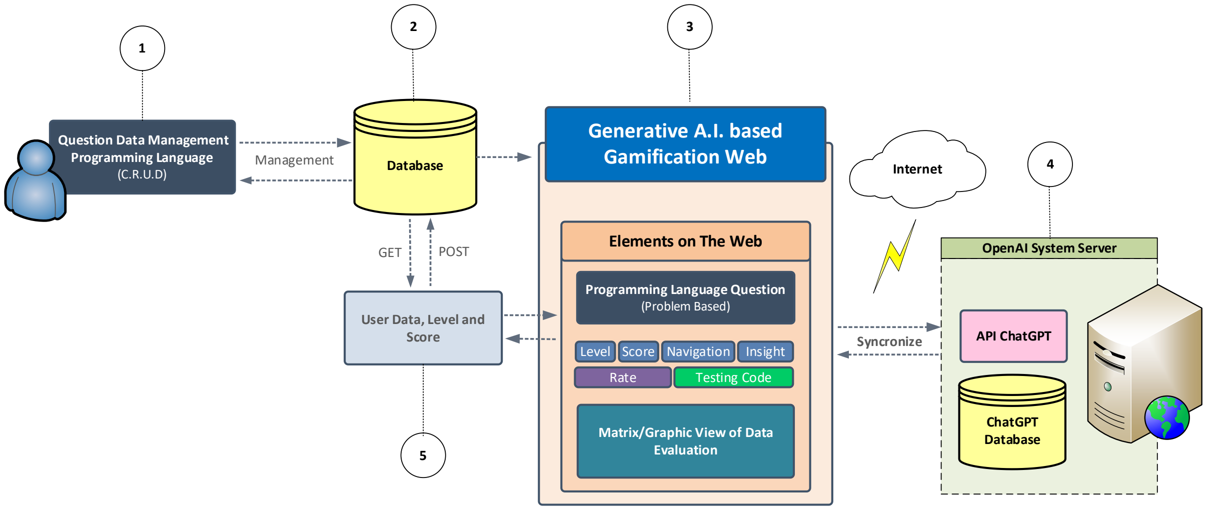

1
2
3
4
5
Question Data Management Programming Language (C.R.U.D)
Management
Database
GET
POST
User Data, Level and Score
Generative A.I. based Gamification Web
Elements on The Web
Programming Language Question (Problem Based)
Level
Score
Navigation
Insight
Rate
Testing Code
Matrix/Graphic View of Data Evaluation
Internet
Syncronize
OpenAI System Server
API ChatGPT
ChatGPT Database

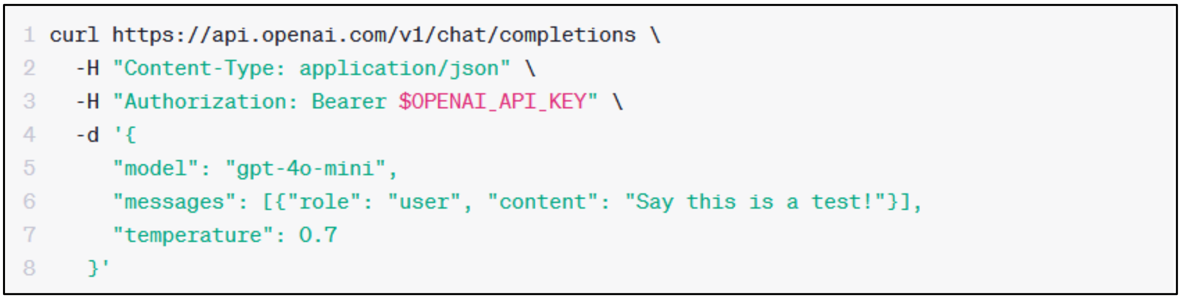
FIGURE 1. Architecture of ChatGPT integrated A.I. based gamification model

Figure 1 shows the proposed architecture model for the development of PBGL system involving artificial intelligence integrated with ChatGPT API. This architecture consists of five main parts that communicate with each other. First, teachers input basic testing materials such as Python, Java, and Web Programming Language, as well as manage assessment criteria, points, and levels for the gamification model of each material through the data input system. Second, the data that has been inputted by the instructor is stored in the system database, which serves as the data management center for the website. Third, student participants of the experimental class access the website, where they can work on questions based on the predetermined difficulty level (level), and receive automatic insights generated by ChatGPT through an integrated API. The website displays information such as the level of difficulty of the material, the score of the correct answer, as well as automatic feedback from the ChatGPT system. Fourth, the developed website system performs synchronization which plays a role in managing communication between the website and the ChatGPT API, ensuring that requests from students are received and processed correctly, and the required data is sent back to the user in the form of insights. Finally, the system also stores students' work results, including scores, levels, and other user data, which can be accessed by teachers and students to view evaluations, result matrices, and feedback, providing a comprehensive picture of the critical thinking skills of experimental class students in gamification and AI-based learning.

6. ARTIFICIAL INTELLIGENCE AND API CHATGPT

A.I. is an intelligent technology capable of interacting with the environment and simulating human intelligence [76] and has adaptation and decision-making capabilities [77] A.I. systems have been applied and developed in society, government, industry, and academia [78-80]. In this research experiment, the utilization of A.I. will be integrated with the ChatGPT API to add information according to the detected material content, so that students can gain additional insight when solving problem-based problems on basic programming materials such as Python, Java and Web Programming Language. The ChatGPT API utilizes Large Language Models (LLMs) that function to understand and generate text naturally. LLMs are trained using large amounts of text data to predict words or phrases in a sentence, allowing students to answer questions, provide information, and carry on conversations in a human-like manner quickly. [81-83]. This integration is expected to improve students' critical thinking skills through the automatic provision of relevant and in-depth information.

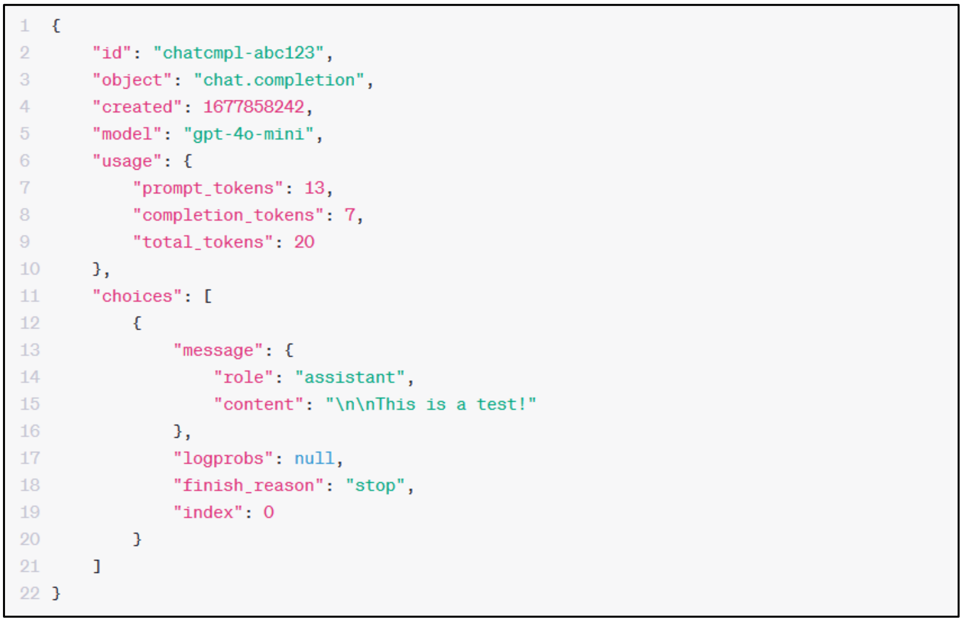
OpenAI provides request functions to enable integration of the ChatGPT API into the client service system to be developed. A brief overview of the request function and the default response that can be used and modified is available on the link page https://platform.openai.com/docs/api-reference/making-requests. The function involves an HTTP POST method with an endpoint URL, authorization header, and parameters in the request body, and returns a JSON object that includes the generated text and additional metadata. The library extensions provided by the ChatGPT API are in the form of node.js, curl and python scripts. Examples of ChatGPT API request and response functions can be seen in figure 2 and figure 3 below.



```
curl https://api.openai.com/v1/chat/completions \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  -d '{
     "model": "gpt-4o-mini",
     "messages": [{"role": "user", "content": "Say this is a test!"}],
     "temperature": 0.7
   }'
```

FIGURE 2. Part of the logic framework performs the request process to the ChatGPT API Endpoint

As explained in Figure 2, this part of the script uses a tool called "curl" to send messages to the virtual assistant from OpenAI. In the first line, it specifies the destination address of the message being sent, which is https://api.openai.com/v1/chat/completions. Then, the next two lines (-H "Content-Type: application/json" and -H "Authorization: Bearer $OPENAI_API_KEY") set the type of information sent (text in JSON format, which is how computers understand data) and enter the secret key (API Key) to ensure only authorized users can use the service. The last line (-d '{...}') contains the message that the user wants to send to the ChatGPT virtual assistant, the selected assistant model (gpt-4o-mini), and the "temperature" setting that determines how creative the answers are. Once these are sent, the ChatGPT virtual assistant will respond according to those instructions.



```
{
    "id": "chatcmpl-abc123",
    "object": "chat.completion",
    "created": 1677858242,
    "model": "gpt-4o-mini",
    "usage": {
        "prompt_tokens": 13,
        "completion_tokens": 7,
        "total_tokens": 20
    },
    "choices": [
        {
            "message": {
                "role": "assistant",
                "content": "\n\nThis is a test!"
            },
            "logprobs": null,
            "finish_reason": "stop",
            "index": 0
        }
    ]
}
```

FIGURE 3. Part of the logic framework of the process of reply response from JSON API ChatGPT from the request result

In Figure 3, the JSON script above is the result of a conversation with the virtual assistant, where "id": "chatcmpl-abc123" is the unique identification number assigned to this conversation, like a serial number that distinguishes one conversation from another. "object": "chat.completion" indicates that this is the result of a completed conversation, meaning that the virtual assistant has provided the answer. The time the conversation occurred is recorded with the number "created": 1677858242, which is how the computer stores time in a format called "epoch time", although to the user this means exactly when the conversation took place. The assistant model used to answer is "model": "gpt-4o-mini", which indicates the version or type of the ChatGPT virtual assistant. In the "usage" section: {...}, there is information about how many words or parts of words (called "tokens") were used: "prompt_tokens": 13 indicates the number of words in the message sent by the user, "completion_tokens": 7 is the number of words used by the assistant to answer, and "total_tokens": 20 is the total number of words. In the "choices" section: [...], it describes the content of the answer given by the ChatGPT assistant, where "message": {...} indicates that the assistant responded with the message "This is a test!". The value "logprobs": null indicates that there is no additional information about the probabilities used by the assistant to select the words in her answer. The reason why the ChatGPT assistant stopped answering is characterized by "finish_reason": "stop", which means that the assistant finished giving the desired answer, and "index": 0 indicates that this is the first (and only) answer in the list of answers given by the assistant in this conversation.

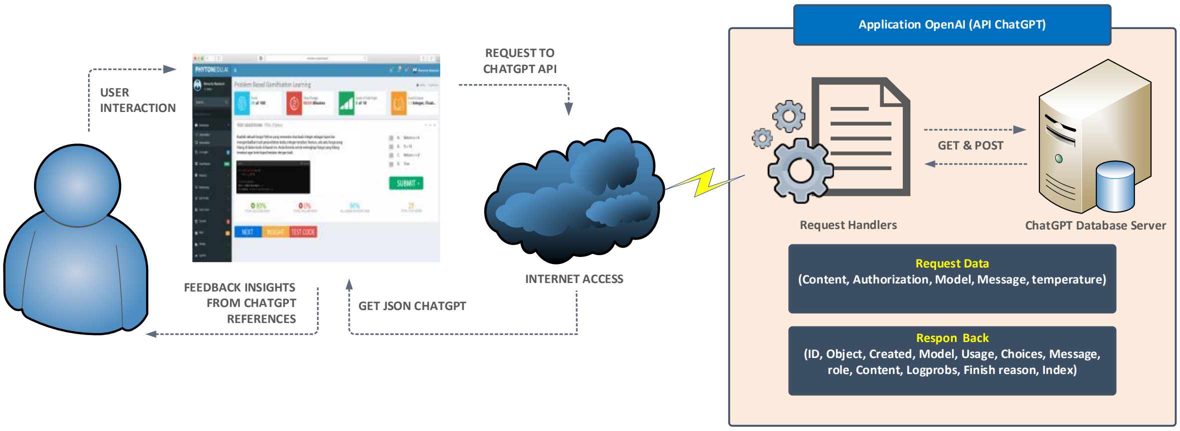

USER INTERACTION
REQUEST TO CHATGPT API
FEEDBACK INSIGHTS FROM CHATGPT REFERENCES
GET JSON CHATGPT
INTERNET ACCESS
Application OpenAI (API ChatGPT)
GET & POST
Request Handlers
ChatGPT Database Server
Request Data
(Content, Authorization, Model, Message, temperature)
Respon Back
(ID, Object, Created, Model, Usage, Choices, Message, role, Content, Logprobs, Finish reason, Index)

FIGURE 4. Architecture of request and response between user and ChatGPT API

The explanation in figure 4 illustrates how the developed system communicates between user requests and the ChatGPT API can work. First, the user interacts with the developed website through the user interface (UI). Here, the user enters the answers of multiple-choice questions into the website system. Once this data is inputted and submitted via the "Submit" button available in the gamification lesson question page, the application sends the information to the ChatGPT database server in JSON format, which is a standardized way of structuring data to be easily understood by computers. Next, the server hosting this database forwards the request in JSON format to the OpenAI server, where the ChatGPT API resides. In this process, the OpenAI server uses a request header containing authorization information and other metadata, which ensures that the request is legitimate and complies with the terms of access. The ChatGPT API then processes the request and generates an answer or response, which is also in JSON format. This response is then sent back to the user's website server, which receives and displays it back through the UI to the user in the form of insights. In addition, the developed website not only displays the results, but also manages the business logic behind the scenes. This includes managing the data stored in the database (DB) for various purposes, such as storing user interaction history so that it can be reused or analyzed in the future.

7. PRE-TEST AND POST-TEST

The pre-test is conducted before the training or material delivery begins with the aim of measuring students' knowledge or skills before receiving training. The pre-test results provide an initial picture of the students' level of understanding of the material to be delivered. In contrast, the post-test is conducted after the training or material delivery is completed to measure students' knowledge or skills after attending the training. The results of the post-test are compared with the results of the pre-test to assess the extent of improvement or change that has occurred in the participants. [84, 85]. In testing the results of initial and final critical thinking skills in student groups (control and experimental groups). The total number of students involved in this research is 520 participants which will be divided randomly and equally, where the control class students are 260 participants and for the experimental class are 260 participants as well, and will be tested pre-test and post-test. The following is a picture of pre-test and post-test testing in the control class and experimental class in figure 5.

FIGURE 5. Control group

In figure 5 is a control class group totaling 260 students. This research process was carried out for 11 months from June 2023 - April 2024, 3 months testing PBGL material involving students from 4 different universities and colleges, namely from East Java Province, East Nusa Tenggara Province (Indonesia), and universities located in the Cova-Lima and Dili districts (Timor Leste). Each student has a duration of 3 months when the researcher provides material about Python, Java and Web Programming Language and this course is a basic and compulsory course (not an elective) for second and third semester students in each University and College which is the object of research. This research was taken because of the problems of each student in understanding complicated programming languages and having a low pass rate, and there is an obligation for each student to pass this compulsory course to be able to take advanced courses in the following semester.

Before the training and testing began, the first step was to administer a pre-test to the control group to assess their basic knowledge and skills in Python, Java, and Web Programming Languages for 3 months. This pre-test serves as a benchmark to understand how well the students have mastered the basic concepts before receiving the training. The control group will undergo training using conventional learning methods, which do not involve gamification elements or artificial intelligence (AI) integration. This method focuses on traditional approaches often used in formal education settings, such as lectures, discussions, and hands-on exercises without the addition of interactive technology. After the training was completed, students were given a post-test identical to the baseline test to measure the extent to which their knowledge and skills had developed during the training. The results of this final test are then compared with the baseline test to determine the effectiveness of conventional learning methods in improving students' ability to understand programming languages. In addition, this comparison also provides insight into how much impact the traditional approach to learning has compared to more modern approaches such as gamification and AI, which may be applied to the experimental group.

FIGURE 6. Experimental group

In Figure 6 is the experimental class group, the total number of students is the same as the control class students which is 260 participants. This research process was carried out for 11 months from June 2023 - April 2024 with students from 4 different universities and colleges, namely from East Java Province, East Nusa Tenggara Province (Indonesia), and Universities located in the Cova-Lima and Dili districts (Timor Leste). Each student has the same duration of 3 months when the researcher provides material on Python, Java and Web Programming Language and this course is a basic and mandatory course (not an elective) for second and third semester students in each University and College that is the object of research. The same problem is the basis of direct examination of these 4 universities because each student in understanding programming languages is complicated and has a low pass rate as well, and there is an obligation for each student to pass this compulsory course to be able to take advanced courses in the following semester.

The experimental group was given a pre-test to measure students' basic knowledge and skills in understanding Python, Java and Web Programming Languages. This test aims to assess the initial level of students' ability in programming before students receive special training. The experimental group then underwent training using the Problem-Based Gamification Learning (PBGL) model, a website integrated with artificial intelligence (AI) through the ChatGPT API. In this method, students get additional information and automatic feedback when solving problem-based problems, which are designed to deepen students' understanding of the basic material concepts of Python, Java and Web Programming Language. The integration of AI enables a more interactive and adaptive learning experience, where students can interact with the system that provides answers, explanations and suggestions according to the difficulties they face. After the training was completed, the same post-test was administered to measure the changes in students' knowledge and skills after going through the training with this advanced PBGL method. The comparison of the pre-test and post-test results in the experimental group will provide insight into the effectiveness of the gamification and AI-based learning approach compared to conventional methods, as well as how much the students' understanding of Python, Java and Web Programming Languages improved as a result of the training.

8. NON-EQUIVALENT CONTROL GROUP DESIGN

This research uses a non-equivalent control group design. This design involves two groups that are not randomly selected, where both groups are given a pre-test to identify initial differences between the experimental group and the control group, here is a research design to test the effectiveness of the PBGL learning model with Python, Java and basic Web Programming Language materials. The following data design non-equivalent control group design can be seen in Table 4.

Table 4. Non-equivalent control group design

Group	Pre-test	Treatment	Post-test
Experiment (A)	O1	X	O2
Control (B)	O3	-	O4

Description:

A : Experimental class with PBGL learning model
B : Control class with conventional method
O1 : Experimental class pre-test
O2 : Post-test of the experimental class
O3 : Control class pre-test
O4 : Control class post-test

X : Treatment/use of PBGL model

Based on the model design in Table 4, this research design will have two analyses conducted. The first analysis aims to compare the initial ability between the experimental group and the control group (O1:O3). To test this difference, the t-test method is used. The expectation is that there is no significant difference between the initial ability of the control group and the experimental group, namely between O1 and O3.

The second analysis aims to test the proposed hypothesis, namely "The application of PBGL model learning will improve critical thinking and problem-solving skills" In this case, the t-test statistical technique for two related samples was used. What is tested is the difference between O2 and O4. If there is a difference where O2 is greater than O1, then learning with the PBGL model has a positive effect. However, if O2 is smaller than O4, then learning with the PBGL model has a negative effect.

IV. DATA ANALYSIS

This research uses various methodological techniques to evaluate the effectiveness of PBGL on critical thinking skills with an A.I. approach integrated with the ChatGPT API. This methodology includes three main tests: content validity ratio testing to assess the suitability of items with the measured domain based on expert judgment, construct validity testing to ensure that the instrument can reflect the measured construct well, and test-retest reliability testing to evaluate the consistency of the instrument in producing stable scores over time. Each of these tests was conducted with appropriate techniques and statistical analysis to ensure the validity and reliability of the data obtained from this study.

9. CONTENT VALIDITY RATIO (CVR) TESTING

Content validity test with CVR is a method used to determine the suitability of items with the measured domain based on the assessment of experts or validators. By using the CVR formula, researchers can find out whether the items developed are in accordance with the objectives to be achieved and whether the measured domain is seen in accordance with the assessment of experts. The CVR formula is expressed as follows.

$$CVR = (2ne/\ n) - 1 \tag{1}$$

Formula Description:

Where *ne* is the number of subject matter experts (SMEs)/penalists/validators who rated an item and *n* is the number of SMEs/penalists/validators who conducted the assessment.

CVR values range from -1 to 1. The higher the CVR value, the better the degree of content validity obtained. An item is considered to have good content validity if the CVR value of the item equals or exceeds the minimum acceptable CVR value. The minimum acceptable CVR value depends on the number of panelists/experts involved in the content validity test.

The use of the CVR method in content validity testing allows researchers to quantitatively measure the extent to which items in an assessment instrument are considered relevant and representative of the domain being measured. This provides a strong basis for ensuring the content validity of the assessment instrument developed. The following are the results of the content validity test with CVR critical thinking and problem-solving skills. The following are the results of the CVR validity test of critical thinking skills by experts in Table 5.

Table 5. CVR validity test results for critical thinking skills (expert)

Topic	Essay Type	NE	CVR	Min Value CVR	Description
Python	Why is initialization important when	8	1	0.75	Valid

	declaration of variables?				
	Create a boolean data type function with an "if" condition example.	8	1	0.75	Valid
	Create a program that performs arithmetic operations on integers.	7	0.75	0.75	Valid
	Create a simple script using assignment operators (=, +=, -=, *=, /=)	7	0.75	0.75	Valid
	Create script logic using the "if-elif-else" function.	8	1	0.75	Valid
	Create scripts using polymorphism for methods from different classes.	7	0.75	0.75	Valid
Java	Explain how to declare and initialize variables in Java with examples.	8	1	0.75	Valid
	Create a program that declares int, float, String, and boolean.	7	0.75	0.75	Valid
	Create a script that uses arithmetic operators, including the modulus operator (%).	7	0.75	0.75	Valid
	Create a program that determines whether a number is positive, negative, or zero using "if-else" in Java.	8	1	0.75	Valid
	Create a script that uses polymorphism to call methods from different classes.	8	1	0.75	Valid
	Create an example of using "private" to protect data in a Java class.	8	1	0.75	Valid
Web Programming Languages	What are the functions of the <head> and <body> elements in an HTML document?	7	0.75	0.75	Valid
	Write HTML examples that use <h1>, <p>, and <a> tags.	8	1	0.75	Valid
	Create a CSS example that organizes the layout of elements with "margins" and "padding".	7	0.75	0.75	Valid
	Create an example of a JavaScript function that is called when the button is clicked.	8	1	0.75	Valid
	Write a JavaScript code example for form validation, checking if the input is empty.	8	1	0.75	Valid
	Write a JavaScript code example that uses "try-catch" to catch and handle errors	7	0.75	0.75	Valid

In Table 5, based on the results of the CVR validity test of critical thinking skills by experts in ensuring the validity of the items in measuring students' critical thinking skills, a validity test using CVR was conducted. This test involves a number of experts (NE) who assess the relevance and accuracy of each item in measuring important aspects of Python, Java, and Web Programming Languages. The CVR obtained indicates the level of agreement among the experts regarding the validity of the items, where a score of 1 indicates full agreement, and a score of 0.75 indicates slight variation in judgment but remains above the minimum required for validity.

The following is a detailed description of each item based on the topic tested, the number of experts involved, and the CVR value obtained. This description aims to provide a deeper understanding of the relevance and effectiveness of each item in the context of teaching and testing students' critical thinking skills.

The following is a detailed description of each item based on the Python topic:

a) Why is initialization important when declaring variables?

It has an NE of 8 and a CVR of 1, indicating that experts agree that this item is highly valid in measuring students' understanding of the importance of variable initialization in programming, where this item effectively measures students' critical ability in understanding the basic concepts of variable declaration and initialization in Python.

b) Create a boolean data type function with an "if" condition example.
It has an NE of 8 and a CVR of 1, which shows that experts agree that the question is very relevant to measure students' critical thinking skills in the use of boolean data types and their application in if conditions. This validity value indicates that this question successfully tests students' ability to apply simple programming logic.

c) Create a program that performs arithmetic operations on integers.
It has an NE of 7 and a CVR of 0.75. Although this CVR score was valid, there was a slight difference of opinion among the experts regarding the relevance of this item. This may be due to variations in understanding of the difficulty level of the item, but it is still recognized as a valid item in measuring students' ability to perform basic arithmetic operations in Python.

d) Create a simple script using assignment operators (=, +=, -=, *=, /=).
It has an NE of 7 and a CVR of 0.75. These scores indicate sufficient validity, but there was not full agreement among the experts. Most likely, this question tested a more specific area in the use of assignment operators, which may have led to variations in relevance ratings.

e) Create script logic using the if-elif-else function.
It has an NE of 8 and a CVR of 1, indicating full agreement among the experts that the question is highly valid in testing the learner's understanding of branching logic using if-elif-else in Python. This high CVR value indicates that the question effectively measures the learner's ability to implement program flow control structures.

f) Create scripts using polymorphism for methods from different classes.
It has an NE of 7 and a CVR of 0.750. While this score is valid, the slight disagreement among experts may be due to the complexity of the polymorphism concept in Python. The question is still recognized as valid to test the learner's ability to apply the principle of polymorphism in a more advanced context.

The following is a detailed description of each item based on Java topics:

a) Explain how to declare and initialize variables in Java with examples.
It has an NE of 8 and a CVR of 1, indicating that the experts agreed that this item is highly valid for testing students' understanding of variable declaration and initialization in Java. This full agreement confirms that the question is able to measure basic understanding in the Java programming language very effectively.

b) Create a program that declares int, float, String, and boolean.
It has an NE of 7 and a CVR of 0.75. Although this CVR score is still valid, slight variations in scoring may arise due to different levels of complexity in understanding different data types in Java. However, this question is still considered valid in measuring the learner's ability to declare variables with different data types.

c) Create a script that uses arithmetic operators, including the modulus operator (%).
It has an NE of 7 and a CVR of 0.75. These scores indicate that this item is valid enough to measure students' understanding of the use of arithmetic operators in Java, including the modulus operator.

However, the small difference in expert ratings suggests that some may see this as a more complex question.

d) Create a program that determines whether a number is positive, negative, or zero using if-else in Java.
It has an NE of 8 and a CVR of 1, indicating full agreement that it is highly relevant and valid for measuring learners' ability to use conditional logic in Java. This CVR value indicates that the question effectively tests understanding of program flow control structures.

e) Create a script that uses polymorphism to call methods from different classes.
It has an NE of 8 and a CVR of 1. This full agreement indicates that the experts agreed that this question is highly valid for measuring students' ability to apply the concept of polymorphism in Java. This high validity indicates that the question effectively measures advanced skills in OOP.

f) Create an example of using private to protect data in a Java class.
It has an NE of 8 and a CVR of 1, indicating full agreement among the experts that this question is highly valid in testing the learner's understanding of encapsulation and the use of private access modifiers to protect data in Java. This question is highly relevant to ensure that students understand the basic principles in OOP.

The following is a detailed description of each item based on the Web Programming Language topic:

a) What are the functions of the <head> and <body> elements in an HTML document?
It has an NE of 7 and a CVR of 0.75. These scores are valid but indicate that there is a slight difference of view among experts on the relevance or complexity of this concept. Nonetheless, it is still considered valid enough to measure a basic understanding of HTML document structure.

b) Write HTML examples that use <h1>, <p>, and <a> tags.
It has an NE of 8 and a CVR of 1. This full agreement indicates that the experts agreed that this item is highly valid in measuring the learner's ability to use common HTML tags such as headings, paragraphs, and links. This is important to ensure a basic understanding of HTML elements.

c) Create a CSS example that organizes the layout of elements with margins and padding.
It has an NE of 7 and a CVR of 0.75. While this score is valid, the slight variation in scoring may be due to variation in understanding of the concept of box models in CSS. The question is still recognized as valid in testing the learner's ability to layout elements using CSS.

d) Create an example of a JavaScript function that is called when the button is clicked.
It has an NE of 8 and a CVR of 1. The experts agreed that this item is highly valid in measuring the learner's ability to use functions and events in JavaScript. This full agreement confirms the relevance of the question in the context of developing interactivity on web pages.

e) Write a JavaScript code example for form validation, checking if the input is empty.
It has an NE of 8 and a CVR of 1, indicating full agreement that it is highly valid in testing the learner's ability to perform client-side form validation using JavaScript. This is important to ensure learners understand the basics of user interaction and input validation.

f) Write a JavaScript code example that uses try-catch to catch and handle errors.
It has an NE of 7 and a CVR of 0.750. While this score is valid, there is a small variation in scoring which may be due to the complexity of understanding the error handling mechanism in JavaScript. However, it is still recognized as valid in testing basic understanding of error handling in web development.

The overall conclusion that the questions had an NE of 8 and a CVR of 1 indicates that the experts agreed that the questions were highly valid in measuring students' critical thinking skills on the topics of Python, Java and Web Programming Languages as listed in Table 5. The experts also agreed that the questions were very effective in assessing students' abilities in these areas. On the other hand, the items with NE of 7 and CVR of 0.75 were also considered valid, although there was a slight difference of opinion among the experts. This suggests that the items are still important, but there may be some experts who have different views on how relevant they are.

10. CONSTRUCT VALIDITY (CV) TESTING

The next validity test is the construct validity test. This test was conducted by 25 students (outside the control and experimental classes) who had the same characteristics and knowledge and materials as the research sample. If the analysis results show that the instrument can reflect the construct well, then construct validity can be considered fulfilled. Otherwise, revisions to the instrument may be necessary. The CV formula is as follows:s

$$r_{xy} = \frac{N\sum xy - _{(\sum x)}(\sum y)}{\sqrt{(N\sum x^2 - (\sum x)^2\,(N\sum y^2 - (\sum y)^2)}} \qquad (2)$$

Where r_{xy} is the correlation coefficient between variable x and variable y, $\sum xy$ is the sum of multiplication between x and y variables, $\sum x^2$ is the sum of squares of x values, $\sum y^2$ is the sum of squares of y values, $(\sum x)^2$ is the sum of the x values then squared, $(\sum y)^2$ is the sum of the y values is then squared.

In this **research**, the CV test was carried out by analyzing the relationship between the items in the instrument and the construct being measured using factor analysis techniques. The results of the construct validity test indicate whether the items correlate with the expected factors and reflect the construct being measured properly. This test is important to ensure that the instruments used in the study have sufficient validity, so that the results obtained can be trusted and relevant to the research objectives. Good construct validity indicates that the instrument is able to provide an accurate description of the critical thinking skills of students measured in the context of the PBGL model integrated with A.I ChatGPT. The following data on the construct validity test of students' critical thinking skills can be seen in Table 6.

Table 6. Construct validity test of critical thinking ability (student)

Topic	Essay Type	r-Count	p-Value	r-Table	Description
Python	Why is initialization important when declaration of variables?	.470	0.018	0.396	Valid
	Create a boolean data type function with an "if" condition example.	.577	0.003	0.396	Valid
	Create a program that performs arithmetic operations on integers.	.688	0.000	0.396	Valid
	Create a simple script using assignment operators (=, +=, -=, *=, /=)	.707	0.000	0.396	Valid
	Create script logic using the "if-elif-else" function.	.565	0.003	0.396	Valid
Java	Create scripts using polymorphism for methods from different classes.	.682	0.000	0.396	Valid
	Explain how to declare and initialize variables in Java with examples.	.424	0.018	0.396	Valid
	Create a program that declares int, float, String, and boolean.	.830	0.003	0.396	Valid

	Create a script that uses arithmetic operators, including the modulus operator (%).	.477	0.000	0.396	Valid
	Create a program that determines whether a number is positive, negative, or zero using "if-else" in Java.	.765	0.000	0.396	Valid
	Create a script that uses polymorphism to call methods from different classes.	.844	0.003	0.396	Valid
	Create an example of using "private" to protect data in a Java class.	.804	0.000	0.396	Valid
Web Programming Languages	What are the functions of the <head> and <body> elements in an HTML document?	.424	0.018	0.396	Valid
	Write HTML examples that use <h1>, <p>, and <a> tags.	.830	0.003	0.396	Valid
	Create a CSS example that organizes the layout of elements with "margins" and "padding".	.477	0.000	0.396	Valid
	Create an example of a JavaScript function that is called when the button is clicked.	.765	0.000	0.396	Valid
	Write a JavaScript code example for form validation, checking if the input is empty.	.844	0.003	0.396	Valid
	Write a JavaScript code example that uses "try-catch" to catch and handle errors	.804	0.000	0.396	Valid

In Table 6, the data is a test of the construct validity of critical thinking skills, to measure the extent to which the items are able to measure students' critical thinking skills in topics covering Python, Java, and Web Programming Languages. This test uses correlation analysis with *r-Count*, *p-Value*, and *r-Table* parameters to determine the validity of each item. The *r-Count* value indicates the strength of the relationship between the item and the construct being measured, while the p-Value determines the statistical significance of the results. The *r-Table* value is used as a reference to assess whether the correlation results are strong enough to be considered valid, where an *r-Count* value greater than the *r-Table* indicates the validity of the item.

The following is a detailed description of each item based on the results of the construct validity test, which includes an analysis of the *r-Count*, *p-Value*, and *r-Table* values, as well as conclusions about the validity of each item in the context of teaching and testing students' critical thinking skills.

The following is a detailed description of each item based on the Python topic:

a) Why is initialization important when declaring variables?

Having an *r-Count* value of .470 and a *p-Value* of 0.018 indicates a moderate positive correlation between this item and the critical thinking construct. This value is greater than the *r-Table* value of .396, indicating that this item is valid in measuring students' understanding of the importance of variable initialization in Python programming.

b) Create a boolean data type function with an "if" condition example.

It has an *r-Count* of .577 and a *p-Value* of .003, indicating a strong positive correlation between this item and the learner's critical thinking skills. This correlation is statistically significant and above the *r-Table* of .396, so this item is considered valid for testing learner understanding of using boolean data types and condition logic.

c) Create a program that performs arithmetic operations on integers.

With an *r-Count* of .688 and a *p-Value* of 0.000, this item shows a very strong positive correlation with the critical thinking construct, which is statistically significant. This value is clearly above the *r-Table* of .396, confirming that this item is valid in measuring the learner's ability to perform arithmetic operations on integers.

d) Create a simple script using assignment operators (=, +=, -=, *=, /=).
The *r-Count* of .707 and *p-Value* of 0.000 indicate a very strong and statistically significant positive correlation between this item and the critical thinking construct. This value is above the *r-Table* of .396, so this item is considered valid in testing students' understanding of the use of assignment operators in Python programming.

e) Create script logic using the if-elif-else function.
With an *r-Count* of .565 and a *p-Value* of .003, this item shows a strong and statistically significant positive correlation, which is above the *r-Table* of .396. This indicates that this item is valid in measuring the learner's ability to apply branching logic using if-elif-else.

f) Create scripts using polymorphism for methods from different classes.
The *r-Count* value of .682 and *p-Value* of 0.000 indicate a very strong and statistically significant positive correlation with the critical thinking construct. This value is above the *r-Table* of .396, indicating that the question is valid in measuring students' understanding of the concept of polymorphism in Python programming.

The following is a detailed description of each item based on Java topics:

a) Explain how to declare and initialize variables in Java with examples.
The *r-Count* of .424 and *p-Value* of .018 indicate a moderate positive correlation between this item and the critical thinking construct, which is above the *r-Table* of .396. This indicates that this item is valid in measuring students' understanding of variable declaration and initialization in Java.

b) Create a program that declares int, float, String, and boolean.
With an *r-Count* of .830 and a *p-Value* of .003, this item shows a very strong positive correlation with the critical thinking construct and is statistically significant. This value is clearly above the *r-Table* of .396, confirming the validity of this item in measuring students' ability to declare variables with different data types.

c) Create a script that uses arithmetic operators, including the modulus operator (%).
The *r-Count* of .477 and *p-Value* of 0.000 indicate a strong and statistically significant positive correlation with the critical thinking construct. This value is above the *r-Table* of .396, indicating that the question is valid in measuring students' understanding of the use of arithmetic operators in Java, including the modulus operator.

d) Create a program that determines whether a number is positive, negative, or zero using if-else in Java.
With an *r-Count* of .765 and a *p-Value* of 0.000, this item shows a very strong and statistically significant positive correlation, which is clearly above the *r-Table of* .396. This indicates that this item is valid in measuring the learner's ability to use conditional logic in Java.

e) Create a script that uses polymorphism to call methods from different classes.
The *r-Count* of .844 and *p-Value of* .003 indicate a very strong and statistically significant positive correlation with the critical thinking construct. This value is above the *r-Table* of .396, indicating that the question is valid in testing students' understanding of polymorphism in Java.

f) Create an example of using private to protect data in a Java class.
The *r-Count* value of .804 and *p-Value* of 0.000 indicate a very strong and statistically significant positive correlation with the critical thinking construct. This value is above the *r-Table* of .396,

indicating that the question is valid in testing the learner's understanding of using the private access modifier to protect data in Java.

The following is a detailed description of each item based on the Web Programming Language topic:

a) What are the functions of the <head> and <body> elements in an HTML document?
The *r-Count* of .424 and *p-Value* of .018 indicate a moderate positive correlation between this item and the critical thinking construct. This value is above the *r-Table* of .396, indicating that this item is valid in measuring the learner's understanding of the function of basic elements in the HTML structure.

b) Write HTML examples that use <h1>, <p>, and <a> tags.
With an *r-Count* of .830 and a *p-Value* of .003, this item shows a very strong and statistically significant positive correlation, which is above the *r-Table of* .396. This indicates that this item is valid in measuring the learner's ability to use common HTML tags.

c) Create a CSS example that organizes the layout of elements with margins and padding.
The *r-Count* of .477 and *p-Value* of 0.000 indicate a strong and statistically significant positive correlation with the critical thinking construct. This value is above the *r-Table* of .396, indicating that this item is valid in measuring the learner's understanding of element layout organization using CSS.

d) Create an example of a JavaScript function that is called when the button is clicked.
With an *r-Count* of .765 and a *p-Value* of 0.000, this item shows a very strong and statistically significant positive correlation with the critical thinking construct. This value is above the *r-Table* of .396, indicating that this item is valid in testing the learner's ability to use functions and events in JavaScript.

e) Write a JavaScript code example for form validation, checking if the input is empty.
The *r-Count* of .844 and *p-Value* of .003 indicate a very strong and statistically significant positive correlation with the critical thinking construct. This value is above the *r-Table* of .396, indicating that this question is valid in measuring the learner's ability to validate forms using JavaScript.

f) Write a JavaScript code example that uses try-catch to catch and handle errors.
The *r-Count* value of .804 and *p-Value* of 0.000 indicate a very strong and statistically significant positive correlation with the critical thinking construct. This value is above the *r-Table* of .396, indicating that the question is valid in measuring the learner's understanding of error handling in JavaScript.

The overall conclusion is that questions that have high *r-count* values, such as *r-counts* of .830, .844, and .804, as well as statistically significant *p-values* ($p < 0.05$), indicate that these questions are highly valid in measuring students' critical thinking skills on topics such as Python, Java, and Web Programming Languages, as listed in Table 6. These questions have very strong correlations with the measured constructs, so they are very effective in assessing students' abilities in these areas.

On the other hand, items with lower *r-Count* values, such as .424 and .470, although still above the *r-Table* of .396, were also considered valid but with lower correlations than items with higher r-Count values. This suggests that the items are still important and relevant, but there is variation in how strongly these items measure the construct of interest. This indicates that while most students may have a good understanding of the topic, there are some areas that may need further emphasis in teaching to ensure all learners can achieve a strong understanding.

Overall, the results of this construct validity test provide a strong basis to support the use of these questions in the evaluation of students' critical thinking skills in Python, Java, and Web Programming Languages. These questions proved to be effective in measuring students' competencies and can be used as reliable evaluation tools in educational contexts.

11. TEST-RETEST RELIABILITY TESTING

After conducting the validity test, the instrument reliability test was also carried out. instrument reliability test in this study using test-retest reliability testing technique. Test-retest reliability is a reliability testing technique by re-testing using the same instrument at different times. stability. To obtain the reliability coefficient through the test-retest approach, it can be done by calculating the linear correlation coefficient between the score on the first measurement (X) and the score on the second measurement (Y). The test-retest reliability coefficient formula is as follows:

$$r_i = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\{N\sum X^2 - (\sum X)^2\}\{N\sum Y^2 - (\sum Y)^2\}}} \qquad (3)$$

Formula Description:

Where r_i is the Test-retest reliability coefficient, N is the number of data pairs, $\sum XY$ is the sum of products X and Y, $\sum X$ is the um of X values, $\sum Y$ is the sum of Y values, $\sum X^2$ is the sum of squares of X values, $\sum Y^2$ is the sum of squares of Y values.

Test-retest reliability of critical thinking ability aims to evaluate whether the instrument produces stable and consistent scores from one testing time to the next testing time, in the context of evaluating critical thinking ability instruments, test-retest reliability shows how consistent the instrument measures an individual's critical thinking ability over time. If students' scores on the initial and retest tests are very similar, it indicates that the instrument has high test-retest reliability in measuring critical thinking.

Calculating test-retest reliability, where students are given critical thinking essays at two different times, the test and retest are given 2 weeks apart (13 meetings over 3 months). Their scores on the two tests were then correlated using the Pearson correlation coefficient statistic, with a high correlation coefficient (0.7 or higher) indicating good test-retest reliability. The results of the critical thinking skills retest can be seen in Table 7.

Table 7. Test-retest reliability of students' critical thinking

		Test	**Retest**
Test	Pearson Correlation	1	.960**
	Sig. (2-tailed)		.000
	N	25	25
Retest	Pearson Correlation	.960**	1
	Sig. (2-tailed)	.000	
	N	25	25

**. Correlation is significant at the 0.01 level (2-tailed).

Based on the results in Table 7, students in this study were asked to complete the same critical thinking test on two different occasions, with a two-week gap between the two tests. The test was designed to measure students' critical thinking skills consistently. The results of these two tests were then compared using Pearson's correlation coefficient, a statistical technique used to measure how strong the relationship is between two sets of data. In this analysis, the Pearson correlation coefficient obtained was .960. This number is very close to 1, indicating that there is a very strong and positive relationship between the first test result and the second test result. That is, students who score high on the first test tend to also score high on the second test, and vice versa. This correlation is also statistically significant, which means that these results are not coincidental, but show a real and reliable pattern.

In addition, the statistical significance of this correlation was tested with a very low *p-value of* 0.000, indicating that the chance of getting a correlation of this magnitude by chance is very small, less than 0.1%

(or less than 1 in 1). In other words, this result is very reassuring in showing that the test has high reliability. Test reliability is important because it shows that the measuring tool used can give consistent results even if it is administered at different times. In a practical context, this means that the critical thinking test used in this study is a tool that can be trusted to measure students' critical thinking skills consistently over time. This gives confidence that the test really measures what it is supposed to measure, and that the results obtained can be used to make accurate conclusions about the critical thinking skills of the students in the PBGL model.

12. ASSUMPTION TEST (HOMOGENEITY)

Before conducting hypothesis testing using the T test, the data needs to first pass assumption testing as a pre-requisite test, which involves testing homogeneity and normality, which are one of the important requirements in the T test. These two tests are very important to ensure that the statistical analysis carried out can provide accurate and reliable results. Once these assumptions are met, then hypothesis testing with the T-test can be done to test for differences between the groups under study.

Homogeneity testing is carried out to ensure that the variances of the two groups of respondents being compared are the same, so that the T test results can be interpreted validly. In this case, the data variance must fulfill the assumption of having the same variance. The statistical test that can be used to test the equality of variance in this study is the Bartlett Levene test. This test is said to be homogeneous if the significance value of the test results is greater than the significance level (0.05). The following are the results of the critical thinking ability homogeneity test can be seen in Table 8.

Table 8. Results of pre-test and post-test homogeneity test of critical thinking skills

Topic	Control Class & Experiment		Levene Statistic	df1	df2	Sig.
Python	Pre-test	Based on Mean	.072	1	518	.776
	Post-test	Based on Mean	.002	1	518	.967
Java	Pre-test	Based on Mean	.025	1	518	.979
	Post-test	Based on Mean	.002	1	518	.967
Web Programming Languages	Pre-test	Based on Mean	.514	1	518	.474
	Post-test	Based on Mean	.002	1	518	.967

Based on the data in Table 8, a homogeneity test was conducted to ensure that the variance between the control and experimental groups on the critical thinking skills pre-test and post-test was the same. This test is important because homogeneity of variance is one of the assumptions that must be met in statistical analysis to compare two groups. If the variance between the control and experimental groups is not homogeneous, the results of further analysis may be less valid. The homogeneity test was conducted using Levene's statistics, which is designed to test the similarity of variance between groups.

The results of this homogeneity test show that for the Python topic, the Levene Statistic value on the pre-test is .072 with a significance value (Sig.) of .776, while on the post-test the Levene Statistic value is .002 with a Sig. value of .967. A Sig. value greater than 0.05 indicates that the variance between the control and experimental groups on the pre-test and post-test is homogeneous, which means there is no significant difference in variance between the two groups.

On the topic of Java, the Levene Statistic value for the pre-test was .025 with a Sig. value of .979, while in the post-test the Levene Statistic value was .002 with a Sig. value of .967. This high Sig. value indicates that the variance between the control and experimental groups was very similar, both before and after the treatment was given, thus confirming that the two groups were homogeneous.

For the topic of Web Programming Languages, the homogeneity test results show that the Levene Statistic value in the pre-test is .514 with a Sig. value of .474, while in the post-test the Levene Statistic value is .002 with a Sig. value of .967. These results also showed homogeneity of variance between the control and experimental groups, both at the pre-test and post-test stages.

Overall, the results of this homogeneity test showed that on all topics tested, both on the pre-test and post-test, the variance between the control and experimental groups was homogeneous. Sig. values that were always greater than 0.05 indicated that there were no significant differences in variance between the groups. Thus, the assumption of homogeneity of variance is met, and the results of further analysis can be considered valid and reliable in measuring students' critical thinking skills. This result ensures that comparisons between the control and experimental groups can be made without bias stemming from differences in variance.

13. ASSUMPTION TEST (NORMALITY)

The results of the normality test in this study were used as a prerequisite before conducting the t-test. Before the data is processed by the t-test, the data must be normally distributed. In this study, the normality test was carried out using the IBM SPSS Statistic 22 program with the Kolmogorov-Smirnov and Shapiro Wilk methods. Data is said to be normally distributed if the significant level is > 0.05, while if the significant level is < 0.05 then the data is not normally distributed. The results of the data normality test analysis of the pre-test and post-test results of critical thinking skills of both groups can be seen in Table 9.

Table 9. Normality test results of pre-test and post-test of critical thinking skills (control and experimental classes)

Topic	Class (Group)		Kolmogorov-Smirnova			Shapiro-Wilk		
			Statistic	df	Sig.	Statistic	Df	Sig.
Python	Pre-test	Control	.031	300	.200*	.994	300	.306
		Experiment	.028	220	.200*	.996	220	.853
	Post-test	Control	.030	300	.200*	.997	300	.813
		Experiment	.054	220	.200*	.991	220	.188
Java	Pre-test	Control	.039	300	.200*	.994	300	.309
		Experiment	.033	220	.200*	.996	220	.860
	Post-test	Control	.030	300	.200*	.997	300	.813
		Experiment	.054	220	.200*	.991	220	.188
Web Programming Languages	Pre-test	Control	.035	300	.200*	.994	300	.313
		Experiment	.029	220	.200*	.996	220	.795
	Post-test	Control	.030	300	.200*	.997	300	.813
		Experiment	.054	220	.200*	.991	220	.188

In Table 9, the results of the normality test of the pre-test and post-test of critical thinking skills in the control and experimental classes for the topics of Python, Java, and Web Programming Languages, displayed in Table 9, show that the data distribution was tested using two statistical methods: Kolmogorov-Smirnov and Shapiro-Wilk. Both tests aim to determine whether the data obtained from the pre-test and post-test are normally distributed, which is one of the basic assumptions in many statistical analyses.

For the Python topic, the Kolmogorov-Smirnov test results on the pre-test showed a statistical value of .031 with a significance value (Sig.) of .200 for the control class, and a statistical value of .028 with a Sig value. .200 for the experimental class. In the post-test, the statistical value for the control class was .030 with a Sig. .200, and for the experimental class was .054 with a Sig value of .200. The Sig. value of .200 in the Kolmogorov-Smirnov test indicates that there is no significant deviation from the normal distribution. The

Shapiro-Wilk test results also support this finding, with the Sig. values all being greater than .05, indicating that the data were normally distributed in both the pre-test and post-test.

The Java topic showed a similar pattern. In the pre-test, the Kolmogorov-Smirnov value for the control class was .039 with a Sig value of .200, and for the experimental class was .033 with a Sig. .200. In the post-test, the statistical value for the control class was .030 with a Sig value. .200, while for the experimental class is .054 with a Sig value of .200. .200. The results of the Shapiro-Wilk test on the pre-test showed a Sig value. .309 for the control class and .860 for the experimental class, which also indicates a normal distribution. In the post-test, the Shapiro-Wilk results showed a Sig value of. .813 for the control class and .188 for the experimental class, which is consistent with the Kolmogorov-Smirnov results in indicating that the data is normally distributed.

For the topic of Web Programming Languages, the Kolmogorov-Smirnov test results on the pre-test showed a statistical value of .035 with a Sig value. .200 for the control class, and .029 with a Sig. .200 for the experimental class. In the post-test, the Kolmogorov-Smirnov value was .030 with a Sig value. .200 for the control class, and .054 with a value of Sig. .200 for the experimental class. The Shapiro-Wilk results were also consistent, with Sig. values in the pre-test of .313 for the control class and .795 for the experimental class, and in the post-test of .813 for the control class and .188 for the experimental class.

Overall, the results of this normality test showed that the data from the pre-test and post-test for all topics and classes tested were normally distributed, as indicated by Sig. values greater than 0.05 in both the Kolmogorov-Smirnov and Shapiro-Wilk tests. This means that the data meets the assumption of normality, which is important for the validity of any further statistical analysis that may be conducted to evaluate the differences between the control and experimental classes in this study.

14. *HYPOTHESIS TEST RESULTS*

Hypothesis testing is a method in statistics used to test the truth of a statement or assumption about population parameters. In other words, hypothesis testing is a procedure that allows decisions to accept or reject statements about the value of a population parameter. The hypothesis test in this study uses an independent sample t-test to determine whether the PBGL model can improve critical thinking skills. The basis for the independent sample t-test decision is if the sig value <0.05 then there is a significant difference between the experimental class and the control class, otherwise if the sig value> 0.05 then there is no difference between the experimental class and the control class. The following will be an independent sample t-test test, as shown in Table 10.

Table 10. Results of independent samples t-test on pre-test and post-test data of critical thinking skills (control and experimental classes)

Topic	Class (Group)		N	Mean	Independent sample Test Sig. (2-tailed)
Phyton	Pre-test	Control	300	49.49	.784
		Experiment	220	50.19	.785
	Post-test	Control	300	51.79	.000
		Experiment	220	61.15	.000
Java	Pre-test	Control	300	50.15	.768
		Experiment	220	50.40	.766
	Post-test	Control	300	51.79	.000
		Experiment	220	61.15	.000
Web Programming Languages	Pre-test	Control	300	50.38	.918
		Experiment	220	50.47	.918
	Post-test	Control	300	51.79	.000

	Experiment	220	61.15	.000

Table 10 shows the results of the independent samples t-test conducted on the pre-test and post-test data of critical thinking skills for the control and experimental classes on the topics of Python, Java, and Web Programming Languages. This t-test is used to determine whether there is a statistically significant difference between the mean (Mean) of the pre-test and post-test results of the two groups.

For the Python topic, the pre-test results showed that the mean score for the control class was 49.49 and for the experimental class was 50.19. The t-test results showed a significance value (Sig.) of .784 for the control class and .785 for the experimental class, which is well above the 0.05 significance limit. This indicates that there was no significant difference between the control and experimental groups at the pre-test stage, so the two groups were considered comparable before the treatment was given. However, in the post-test, there was a significant difference between the two groups, with a mean score of 51.79 for the control class and 61.15 for the experimental class. The significance value of 0.000 shows a highly significant difference between the two groups after the treatment was given, which indicates that the experimental treatment had a real impact on improving critical thinking skills on Python topics.

For the Java topic, the pre-test results showed a mean score of 50.15 for the control class and 50.40 for the experimental class, with significance values of .768 and .766 respectively. Just like in the Python topic, this value shows no significant difference between the control and experimental groups at the pre-test stage. However, in the post-test, there was a significant difference between the control and experimental groups, with a mean score of 51.79 for the control class and 61.15 for the experimental class, and a significance value of 0.000. This indicates that the experimental treatment on the Java topic also had a significant impact on improving critical thinking skills.

For the Web Programming Languages topic, the pre-test results showed a mean score of 50.38 for the control class and 50.47 for the experimental class, with a significance value of .918 for both groups. This indicates that there was no significant difference between the control and experimental groups at the pre-test stage. However, in the post-test, there was a significant difference with a mean score of 51.79 for the control class and 61.15 for the experimental class, and a significance value of .000. These results indicate that the experimental treatment also had a significant impact on critical thinking skills on the topic of Web Programming Languages.

Overall, the results of the independent samples t-test showed that there were no significant differences between the control and experimental groups at the pre-test stage in all topics tested, indicating that the two groups were comparable before treatment. However, significant differences were found in the post-test across all topics, indicating that the treatment provided in the experiment was effective in improving students' critical thinking skills on the topics of Python, Java, and Web Programming Languages. The visualization of the pre-test and post-test data results before and after treatment from the mean results can be seen in Figure 7.

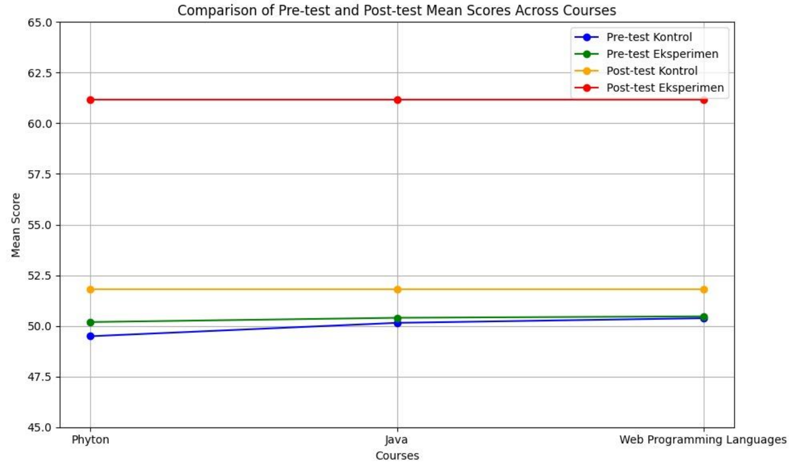


FIGURE 7. PBGL material testing room view

Based on the visualization of Figure 7, it can be concluded that there is a significant difference between the experimental group and the control group. With regard to the hypothesis test in this study, namely:

1. **H0** : The use of PBGL learning model cannot improve critical thinking skills.
2. **H1** : The use of PBGL learning model can improve critical thinking skills

So that the initial hypothesis (**H0)** is rejected. Thus, the use of PBGL learning model with A.I-based gamification approach integrated with ChatGPT API can significantly improve critical thinking skills compared to conventional learning methods. The following are the results of the pre-test and post-test critical thinking skills gain score test in Table 11.

Table 11. Results of pre-test and post-test gain score test of critical thinking skills

Gain Score	Class (Group)	N	Mean
Gain_score	Control	300	8.20
Phyton	Experiment	220	20.75
Gain_score	Control	300	7.86
Java	Experiment	220	25.58
Gain_score	Control	300	8.58
Web Programming Languages	Experiment	220	23.74

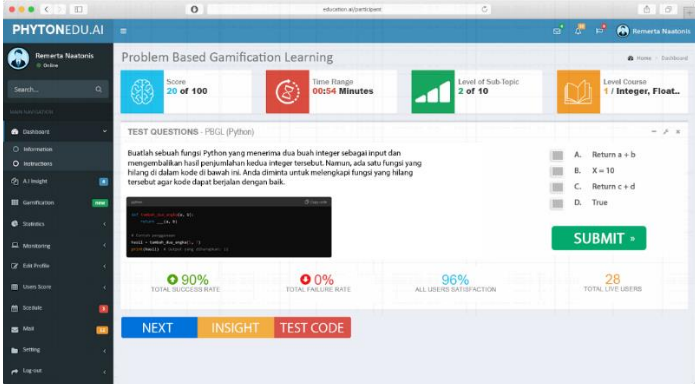
Table 11 shows the results of the gain score test calculated from the difference between the pre-test and post-test scores of critical thinking skills in the control and experimental classes for the topics of Python, Java, and Web Programming Languages. This gain score is used to evaluate how much critical thinking skills improved after treatment in the experimental group compared to the control group.

On the topic of Python, the average gain score for the control class was 8.20, while for the experimental class it was 20.75. This significant difference between the two groups indicates that students in the experimental group experienced a much greater improvement in critical thinking skills after receiving the treatment compared to the control group. For the Java topic, the average gain score for the control class was 7.86, while that for the experimental class was 25.58. This also shows a significant difference in the improvement of critical thinking skills, where the experimental group showed a much greater improvement than the control group after the treatment was given. Finally, on the topic of Web Programming Languages, the average gain score for the control class was 8.58, while for the experimental class was 23.74. These results are consistent with the previous findings, which showed that the experimental group experienced a greater improvement in critical thinking skills than the control group.

Overall, the results of this gain score test show that the treatment provided in the experimental group significantly improved students' critical thinking skills on all topics tested compared to the control group. The large difference in gain score between the two groups confirms the effectiveness of the treatment in improving critical thinking skills in various programming areas. Therefore, it can be concluded that the use of the PBGL model in the experimental group has a positive influence on the increase in scores when compared to the control group. The experimental group had a greater/higher score increase than the control group, and this difference was statistically significant. These results indicate that the learning model used in the experimental group, namely PBGL with an A.I. approach integrated with the ChatGPT API can improve critical thinking skills.

15. *GAMIFICATION WEBSITE (INTEGRATED A.I CHATGPT API)*

Utilizing the ChatGPT API function can provide accurate responses for various types of knowledge-based assessment features, such as multiple-choice questions, SAQ (short answer questions), SEQ (structured essay questions), true/false, and fill in the blank. However, its capabilities are limited to text-based questions only [86]. Similar features to the service developed as a test of critical thinking skills, where programmingedu.ai is an intelligent website integrated with the ChatGPT API to test experimental classes in solving basic programming material questions such as Python, Java and Web Programming Language based on gamification or PBGL. Students will receive basic programming problems presented in the form of problems that vary in model, where each problem is designed to challenge the problem-solving skills of each experimental class student. The system utilizes the ChatGPT API to provide automatic feedback and adaptive learning support, ensuring each student gets the insight needed to understand the basic concepts of programming language upon successfully answering the correct answer. Through a gamification approach, students are encouraged to continue learning and completing tasks in a more engaging and interactive way. Each answer submitted by students will be evaluated automatically, and they will get a score based on the difficulty of the question and the accuracy of the answer. This score can then be used to unlock new challenges or levels, adding an element of competition and motivation to the learning process. In addition, student performance statistics will be recorded for further analysis, allowing for a comprehensive evaluation of the effectiveness of the PBGL method. The website aims to increase student engagement and learning effectiveness through the integration of advanced technology and innovative pedagogical approaches. The appearance of programmingedu.ai website can be seen in Figure 8.



Problem Based Gamification Learning

Score 20 of 100

Time Range 00:54 Minutes

Level of Sub-Topic 2 of 10

Level Course 1 / Integer, Float..

TEST QUESTIONS - PBGL (Python)

Buatlah sebuah fungsi Python yang menerima dua buah integer sebagai input dan mengembalikan hasil penjumlahan kedua integer tersebut. Namun, ada satu fungsi yang hilang di dalam kode di bawah ini. Anda diminta untuk melengkapi fungsi yang hilang tersebut agar kode dapat berjalan dengan baik.

```
def tambah_dua_angka(a, b):
    return __(a, b)

# Contoh penggunaan
hasil = tambah_dua_angka([illegible])
print(hasil)  # Output yang diharapkan: [illegible]
```

- A. Return a + b
- B. X = 10
- C. Return c + d
- D. True

90% TOTAL SUCCESS RATE

0% TOTAL FAILURE RATE

96% ALL USERS SATISFACTION

28 TOTAL LIVE USERS

FIGURE 8. PBGL material testing room view

In Figure 8, the display and main menu available in the programmingedu.ai trial service are described, which includes experimental student data information when logging in and starting to answer questions. The "Score" information menu displays the results obtained by students regarding success or failure in answering questions, while "Time Range" shows the average time taken by students to complete the given question. "Level of Sub-topic" indicates the type of question topic that students are working on, and "Course Level" indicates the level of the question being worked on. The "Test Question" menu presents the form of questions read and answered in the system with various types (essay or multiple choice), and displays statistics or percentages of students' ability to answer questions in real-time. Navigation buttons include the "Next" button to proceed to the next question, the "Insight" button to see the new understanding generated by the ChatGPT API in real-time if the student's answer is correct (the "Insight" button will be disabled if the answer is wrong), and the "Test Code" button that allows students to test the code to ensure the correctness of the answers given. In this research, the frame function on "Test Code" uses the CodeMirror library to debug Python, Java and Web scripts directly.

V. DISCUSSION

The implementation of research on the implementation of PBGL integrated with A.I through the ChatGPT API showed a significant increase in students' critical thinking skills. [87] in various programming topics, including Python, Java, and Web Programming Languages. This study used a non-equivalent control group design with pre-test and post-test evaluation to compare the experimental group using PBGL and A.I integration with the control group using conventional learning methods.

The evaluation results of A.I-based PBGL integrated with ChatGPT API showed a significant increase in critical thinking skills in the experimental group in all topics tested. [88]. For the Python topic, there was an average increase of 20.75 points compared to 8.20 points in the control group, showing an average difference of 12.55 points. Meanwhile, on the Java topic, the experimental group experienced an average increase of 25.58 points, much higher than the 7.86 points in the control group, showing an average difference of 17.72 points. On the topic of Web Programming Languages, the experimental group also showed a significant improvement with an average gain score of 23.74 points compared to 8.58 points in the control group, showing an average difference of 15.16 points.

Rigorous research methodology, including CVR, CV, test-retest reliability, homogeneity, normality and T-test testing ensured the robustness of the findings. CVR values were more than 0.75 for all essay questions (Python, Java, Web Programming Language), indicating that the questions were relevant to the research objectives. Construct validity was also confirmed through CV testing, which showed that the instrument was able to measure the desired students' critical thinking skills with significant correlations. In addition, the reliability of the instrument was tested using the Test-retest Reliability method, and the results showed excellent consistency with a Pearson correlation coefficient of .960. Based on the homogeneity test for the two groups of data, it was found that the data were homogeneous with a p-value> 0.05 and the normality test results found that the data were normally distributed with a p-value> 0.05. Meanwhile, hypothesis testing using the t-test found that the data in the pre-test showed that both the control and experimental groups had the same initial ability in critical thinking, with a significance value of .480> 0.05, which indicated that there was no significant difference between the two. However, after the intervention, the post-test showed a significance value of 0.000 < 0.05, indicating a significant difference in the improvement of students' critical thinking skills. The gain score test revealed that the experimental group experienced a much greater increase in scores than the control group, where on the Python programming language learning had an average difference of 20.75 points compared to 8.20 points in the control group, indicating an average difference of 12.55 points. Meanwhile, on the topic of Java, the experimental group experienced an average increase of 25.58 points, much higher than the 7.86 points in the control group, showing an average difference of 17.72 points. On the topic of Web Programming Languages, the experimental group also showed a significant improvement with an average gain score of 23.74 points compared to 8.58 points in the control group, showing an average difference of 15.16 points.

Thus, the integration of A.I. through ChatGPT provides automated and relevant feedback and additional insights that enrich the learning experience and support deeper understanding across all three programming topics. [89, 90]. This research confirms the potential of A.I-enhanced gamification in education to improve critical thinking skills not only in one programming topic but also in other topics such as Java and Web Programming Languages. These results offer an adaptive and engaging model of future curriculum design. [91], which can be widely applied to improve critical thinking skills in programming education and other fields. [92].

VI. CONCLUSION

The results of this study show that the PBGL model integrated with artificial intelligence through ChatGPT API significantly improves learners' critical thinking skills in learning basic programming such as Python, Java, and Web Programming Language. The group of learners who followed the experimental method showed a significantly higher increase in critical thinking scores compared to the control group, which did not use the PBGL approach. These results prove that the A.I-enhanced gamified learning approach is effective in improving learner engagement and critical thinking skills. The findings of the statistical evaluation include content validity confirmed through CVR values >= 0.75 for all essay questions, while construct validity is evidenced by significant correlations between the instrument items and the critical thinking construct. Instrument reliability was also assured with a Pearson correlation coefficient of .960 on the test-retest reliability test. Data normality and homogeneity tests showed that the data were normally distributed and homogeneous, so the t-test could be used to test the hypothesis. The t-test results showed that there was no significant difference between the experimental and control groups in the initial critical thinking ability (pre-test). However, after the treatment, there was a significant difference in the final critical thinking ability (post-test), where the experimental group experienced a significant increase compared to the control group. This is reinforced by the results of the gain score test which shows a much greater difference in score improvement in the experimental group for all learning topics (Python, Java, and Web Programming Language). Thus, it can be concluded that the learning method applied to the experimental group is effective in improving significant critical thinking skills. Thus, the initial hypothesis (H0) stating that the PBGL model does not improve critical thinking skills is rejected or not proven, and the alternative hypothesis (H1) is accepted, confirming the effectiveness of using this A.I-based PBGL model.

These findings highlight the great potential in integrating A.I into gamification-based educational models to provide personalized feedback according to individual needs. Future research needs to explore the long-term impact of using A.I. and address technical challenges that may arise, so that the integration of A.I. in various educational contexts can be optimized.

The limitation of this research is to evaluate the effectiveness of students' critical thinking skills by using an artificial intelligence (A.I.) based approach integrated with the ChatGPT API. This research is focused on the basic courses of Python, Java, and Web Programming Language. With the help of an adaptive ChatGPT database, this research seeks to provide insight and understanding of additional material automatically to students. The results of the research are expected to reveal the extent to which the use of ChatGPT contributes to improving students' critical thinking skills in these courses and provide more personalized insights.

VII. IMPLICATION

The practical implication of this research is that teachers and policy makers in the field of education need to consider applying the PBGL learning model to improve students' critical thinking skills. By utilizing gamification elements, such as challenges (levels), competitions, and rewards, points, will make the learning process more interesting and interactive. Teachers need to consider the integration of PBGL in curriculum design and teaching methods. The use of A.I-based gamification with integrated ChatGPT API in learning not only increases student motivation and engagement but also enables better development of critical thinking skills, as well as providing a more personalized and adaptive experience for students. In addition, the importance of student learning interest needs to be considered in curriculum design and learning strategies. This factor will certainly contribute to the effectiveness of learning and the development of critical thinking skills.

Funding Statement

This research is supported by the Government of Indonesia, namely the Education Fund Management Agency (LPDP) and the Higher Education Financing Center (BPPT) as funders/sponsors in this research.

Author Contributions

Conceptualization: Remerta Noni Naatonis, Rusijono, and Miftakhul Jannah; Methodology: Remerta Noni Naatonis, Rusijono, and Miftakhul Jannah; Software: Edwin Ariesto Umbu Malahina; Validation: Remerta Noni Naatonis, and Edwin Ariesto Umbu Malahina; Formal Analysis: Remerta Noni Naatonis.; Investigation: Remerta Noni Naatonis, and Edwin Ariesto Umbu Malahina; Data Curation: Remerta Noni Naatonis., and Edwin Ariesto Umbu Malahina.; Writing Original Draft Preparation: Remerta Noni Naatonis., and Edwin Ariesto Umbu Malahina.; Writing Review and Editing: Rusijono, and Miftakhul Jannah; Visualization: Remerta Noni Naatonis, Rusijono, and Miftakhul Jannah; All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

The authors declare that they have no competing financial interests or personal relationships that could influence the work reported in this paper.

Data Availability Statement

Data are available from the authors upon request.

Acknowledgement

The researcher would like to express his sincere and deep gratitude during the preparation of this article which was prepared based on the results of research conducted at the expense of budget funds from the Government of Indonesia, namely Lembaga Pengelola Dana Pendidikan (LPDP) and Balai Pembiayaan Pendidikan Tinggi (BPPT) as funders/sponsors in this research.

REFERENCES

1. Čubela, D., Rossner, A., & Neis, P. (**2023**). Using problem-based learning and gamification as a catalyst for student engagement in data-driven engineering education: A report. *Educ Sci (Basel)*, *13*(12), 1223–1235.
2. Daba, J. B. R., Rosmansyah, Y., & Dabarsyah, B. (**2019**). Problem based learning using gamification: A systematic literature review. In *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)* (pp. 125-130). IEEE.
3. Barua, A. M., & Bharali, S. S. (**2023**). Gamification and its challenges in e-learning: A case study of computer science learners at KKHSOU. *Asian Association of Open Universities Journal*, *18*(3), 233-245.
4. Taesotikul, T., Chinpaisal, C., & Nawanopparatsakul, S. (**2021**). Kahoot! gamification improves learning outcomes in problem-based learning classroom. In *2021 3rd International Conference on Modern Educational Technology* (pp. 125-129). ACM.
5. Kladchuen, R., & Srisomphan, J. (**2021**). The synthesis of a model of problem-based learning with the gamification concept to enhance the problem-solving skills for high vocational certificate. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(14), 4-21.

6. Poonsawad, A., Srisomphan, J., & Sanrach, C. (**2022**). Synthesis of problem-based interactive digital storytelling learning model under gamification environment promotes students' problem-solving skills. *International Journal of Emerging Technologies in Learning (iJET)*, *17*(05), 103-119.
7. Imran, H. (**2023**). An empirical investigation of the different levels of gamification in an introductory programming course. *Journal of Educational Computing Research*, *61*(4), 847-874.
8. Gejandran, P., & Abdullah, N. (**2024**). Gamification in e-learning: A systematic review of benefits, challenges, and future possibilities. *Journal of Logistics, Informatics and Service Science*, *11*(2), 84-104.
9. Swacha, J. (**2022**). Topic evolution in the research on educational gamification. *Educ Sci (Basel)*, *12*(10), 640-653.
10. Vijay, M., & Jayan, J. P. (**2023**). Evolution of research on adaptive gamification: A bibliometric study. In *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)* (pp. 1062-1070). IEEE.
11. Dichev, C., & Dicheva, D. (**2017**). Gamifying education: What is known, what is believed and what remains uncertain: A critical review. *International Journal of Educational Technology in Higher Education*, *14*(1), 9-24.
12. Oliveira, W., Hamari, J., Shi, L., Toda, A. M., Rodrigues, L., Palomino, P. T., & Isotani, S. (**2023**). Tailored gamification in education: A literature review and future agenda. *Educ Inf Technol (Dordr)*, *28*(1), 373-406.
13. Koravuna, S., & Surepally, U. K. (**2020**). Educational gamification and artificial intelligence for promoting digital literacy. In *Proceedings of the 2nd International Conference on Intelligent and Innovative Computing Applications* (pp. 1-6). ACM.
14. Bennani, S., Maalel, A., & Ben Ghezala, H. (**2022**). Adaptive gamification in e-learning: A literature review and future challenges. *Computer Applications in Engineering Education*, *30*(2), 628-642.
15. Elbanna, S., & Armstrong, L. (**2024**). Exploring the integration of ChatGPT in education: Adapting for the future. *Management & Sustainability: An Arab Review*, *3*(1), 16-29.
16. Rejeb, A., Rejeb, K., Appolloni, A., Treiblmaier, H., & Iranmanesh, M. (**2024**). Exploring the impact of ChatGPT on education: A web mining and machine learning approach. *The International Journal of Management Education*, *22*(1), 1-14.
17. AlAli, R., & Wardat, Y. (**2024**). Enhancing classroom learning: ChatGPT's integration and educational challenges. *International Journal of Religion*, *5*(6), 971-985.
18. Castonguay, A., Farthing, P., Davies, S., Vogelsang, L., Kleib, M., Risling, T., & Green, N. (**2023**). Revolutionizing nursing education through AI integration: A reflection on the disruptive impact of ChatGPT. *Nurse Educ Today*, *129*(10), 105916–105929.
19. Angelelli, C. V., Ribeiro, G. M. de C., Severino, M. R., Johnstone, E., Borzenkova, G., & da Silva, D. C. O. (**2023**). Developing critical thinking skills through gamification. *Think Skills Creat*, *49*(1), 101354–101366.
20. Rusandi, M. A., Ahman, Saripah, I., Khairun, D. Y., & Mutmainnah. (**2023**). No worries with ChatGPT: Building bridges between artificial intelligence and education with critical thinking soft skills. *J Public Health (Bangkok)*, *45*(3), e602–e603.
21. Supnoon, A., & Chonchaiya, R. (**2024**). A study on the development of eleventh grade students' critical thinking skills and self-efficacy using active learning pedagogy with gamification. *International Journal of Education and Practice*, *12*(2), 447-466.
22. Heliawati, L., Lidiawati, L., & Pursitasari, I. D. (**2022**). Articulate Storyline 3 multimedia based on gamification to improve critical thinking skills and self-regulated learning. *International Journal of Evaluation and Research in Education (IJERE)*, *11*(3), 1435–1444.
23. Jodoi, K., Takenaka, N., Uchida, S., Nakagawa, S., & Inoue, N. (**2021**). Developing an active-learning app to improve critical thinking: Item selection and gamification effects. *Heliyon*, *7*(11), 8256–8263.
24. Tzelepi, M., Makri, K., Petroulis, I., Moundridou, M., & Papanikolaou, K. (**2020**). Gamification in online discussions: How do game elements affect critical thinking? In *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)* (pp. 92-94). IEEE.
25. Huang, L. Y., & Yeh, Y. C. (**2017**). Meaningful gamification for journalism students to enhance their critical thinking skills. *International Journal of Game-Based Learning*, *7*(2), 47-62.
26. Mahdi, H. S., & Sahari, Y. M. (**2024**). Exploring the relationship between critical thinking, attitude, and anxiety in shaping the adoption of artificial intelligence in translation among Saudi translators. *Journal of Pedagogical Research*, *8*(2), 81-94.
27. Nappi, C., & Cuocolo, A. (**2020**). The machine learning approach: Artificial intelligence is coming to support critical clinical thinking. *Journal of Nuclear Cardiology*, *27*(1), 156-158.
28. Kuchyn, I. L., Lymar, L. V., Bielka, K. Y., Storozhuk, K. V., & Kolomiiets, T. V. (**2024**). New training, new attitudes: Non-clinical components in Ukrainian medical PhD training (regarding critical thinking, academic integrity, and artificial intelligence use). *Wiadomości Lekarskie*, *77*(4), 665-669.
29. Jia, X. H., & Tu, J. C. (**2024**). Towards a new conceptual model of AI-enhanced learning for college students: The roles of artificial intelligence capabilities, general self-efficacy, learning motivation, and critical thinking awareness. *Systems*, *12*(3), 74.
30. Kuhn, D. (**2019**). Critical thinking as discourse. *Hum Dev*, *62*(3), 146-164.
31. Prokop-Dorner, A., Piłat-Kobla, A., Ślusarczyk, M., Świątkiewicz-Mośny, M., Ożegalska-Łukasik, N., Potysz-Rzyman, A., Zarychta, M., Juszczyk, A., Kondyjowska, D., Magiera, A., Maraj, M., Storman, D., Warzecha, S., Węglarz, P., Wojtaszek-Główka, M., Żabicka, W., & Bała, M. M. (**2024**). Teaching methods for critical thinking in health education of children up to high school: A scoping review. *PLoS One*, *19*(7), e0307094–e0307094.
32. Shafieieh, M., Ozturen, A., Rezapouraghdam, H., & Karatepe, O. M. (**2024**). Does critical thinking mediate the relationship between sustainability knowledge and tourism students' ability to make sustainable decisions? *Sustainability*, *16*(13), 5655–5671.

33. Butler, H. A. (**2024**). Predicting everyday critical thinking: A review of critical thinking assessments. *J Intell*, *12*(2), 16-29.
34. 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), 39-66.
35. Orji, C. T., & Ogbuanya, T. C. (**2022**). Mediating roles of ability beliefs and intrinsic motivation in PBL and engagement in practical skills relations among electrical/electronic education undergraduate. *Innovations in Education and Teaching International*, *59*(3), 326-336.
36. Farikhatul Mutma'innah, & Hamimi, E. (**2024**). Development of PBL-based GLOWASEA (Global Warming on the Sea) educational media to train critical thinking skills on the topic of global warming. *International Journal of Interactive Mobile Technologies (iJIM)*, *18*(09), 117-140.
37. Fajari, L. E. W., Sarwanto, & Chumdari. (**2020**). Improving elementary school's critical thinking skills through three different PBL-assisted learning media viewed from learning styles. *Journal of E-Learning and Knowledge Society*, *16*(1), 55-64.
38. Crespí, P., García-Ramos, J. M., & Queiruga-Dios, M. (**2022**). Project-based learning (PBL) and its impact on the development of interpersonal competences in higher education. *Journal of New Approaches in Educational Research*, *11*(2), 259-276.
39. Azer, S. A. (**2004**). Twelve tips for becoming a student in a PBL course: Twelve tips for successful group discussion. *Med Teach*, *26*(1), 12-15.
40. Loyens, S. M. M., Magda, J., & Rikers, R. M. J. P. (**2008**). Self-directed learning in problem-based learning and its relationships with self-regulated learning. *Educ Psychol Rev*, *20*(4), 411-427.
41. Afzaal, M., Nouri, J., Zia, A., Papapetrou, P., Fors, U., Wu, Y., Li, X., & Weegar, R. (**2021**). Explainable AI for data-driven feedback and intelligent action recommendations to support students self-regulation. *Front Artif Intell*, *4*(1), 1-20.
42. Ruiz-Rojas, L. I., Salvador-Ullauri, L., & Acosta-Vargas, P. (**2024**). Collaborative working and critical thinking: Adoption of generative artificial intelligence tools in higher education. *Sustainability*, *16*(13), 5367–5390.
43. Wong, R. S. Y. (**2024**). ChatGPT in medical education: Promoting learning or killing critical thinking? *Education in Medicine Journal*, *16*(2), 177-183.
44. Dahalan, F., Alias, N., & Shaharom, M. S. N. (**2024**). Gamification and game based learning for vocational education and training: A systematic literature review. *Educ Inf Technol (Dordr)*, *29*(2), 1279–1317.
45. Zeybek, N., & Saygı, E. (**2024**). Gamification in education: Why, where, when, and how—a systematic review. *Games Cult*, *19*(2), 237-264.
46. Dehghanzadeh, H., Farrokhnia, M., Dehghanzadeh, H., Taghipour, K., & Noroozi, O. (**2024**). Using gamification to support learning in K-12 education: A systematic literature review. *British Journal of Educational Technology*, *55*(1), 34-70.
47. Castellano, M. S., Contreras-McKay, I., Neyem, A., Farfán, E., Inzunza, O., Ottone, N. E., del Sol, M., Alario-Hoyos, C., Alvarado, M. S., & Tubbs, R. S. (**2024**). Empowering human anatomy education through gamification and artificial intelligence: An innovative approach to knowledge appropriation. *Clinical Anatomy*, *37*(1), 12-24.
48. Bachiri, Y. A., Mouncif, H., & Bouikhalene, B. (**2023**). Artificial Intelligence empowers gamification: Optimizing student engagement and learning outcomes in e-learning and MOOCs. *International Journal of Engineering Pedagogy (iJEP)*, *13*(8), 4-19.
49. Freire-Palacios, V., Jaramillo-Galarza, K., Quito-Calle, J., & Orozco-Cantos, L. (**2023**). La inteligencia artificial en la gamificación para promover la salud mental de los estudiantes universitarios: una revisión de alcance. *Salud, Ciencia y Tecnología*, *3*(1), 639-656.
50. Alahmari, M., Jdaitawi, M. T., Rasheed, A., Abduljawad, R., Hussein, E., Alzahrani, M., & Awad, N. (**2023**). Trends and gaps in empirical research on gamification in science education: A systematic review of the literature. *Contemp Educ Technol*, *15*(3), ep431–ep447.
51. Smiderle, R., Rigo, S. J., Marques, L. B., Peçanha de Miranda Coelho, J. A., & Jaques, P. A. (**2020**). The impact of gamification on students' learning, engagement and behavior based on their personality traits. *Smart Learning Environments*, *7*(1), 3-15.
52. Chan, K. S., & Zary, N. (**2019**). Applications and challenges of implementing artificial intelligence in medical education: An integrative review. *JMIR Medical Education*, *5*(1), e13930–e13945.
53. Bachiri, Y. A., & Mouncif, H. (**2023**). Artificial Intelligence system in aid of pedagogical engineering for knowledge assessment on MOOC platforms: Open EdX and Moodle. *International Journal of Emerging Technologies in Learning (iJET)*, *18*(05), 144-160.
54. Naidu, K., & Sevnarayan, K. (**2023**). ChatGPT: An ever-increasing encroachment of artificial intelligence in online assessment in distance education. *Online Journal of Communication and Media Technologies*, *13*(3), e202336–e202347.
55. Tsai, Y. C. (**2023**). Empowering learner-centered instruction: Integrating ChatGPT Python API and Tinker learning for enhanced creativity and problem-solving skills. In *Innovative Technologies and Learning* (1st ed., pp. 531-541). Cham: Springer.
56. Резаев, А. В., & Трегубова, Н. Д. (**2023**). От социологии алгоритмов к социальной аналитике искусственной социальности: анализ кейсов API и ChatGPT. *The Monitoring of Public Opinion: Economic & Social Changes*, *3*(3), 3-22.
57. Babu, S. S., & Moorthy, A. D. (**2024**). Application of artificial intelligence in adaptation of gamification in education: A literature review. *Computer Applications in Engineering Education*, *32*(1), e22683–e22698.
58. Yu, P., & Wang, S. (**2024**). An examination and analysis of the integration of artificial intelligence and gamification in the pedagogy of Chinese higher education. In *Engaged Learning and Innovative Teaching in Higher Education* (1st ed., pp. 29-46). Singapore: Springer.

59. Modi, T., & Gochhait, S. (**2023**). Impact of artificial intelligence on gamification: Current applications. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)* (pp. 287-290). Uttarakhand: IEEE.
60. van den Berg, G., & du Plessis, E. (**2023**). ChatGPT and generative AI: Possibilities for its contribution to lesson planning, critical thinking, and openness in teacher education. *Educational Sciences (Basel)*, *13*(10), 998–1010.
61. Alarcón-López, C., Krütli, P., & Gillet, D. (**2024**). Assessing ChatGPT's influence on critical thinking in sustainability-oriented activities. In *2024 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1-10). Kos Island, Greece: IEEE.
62. Pishtari, G., Sarmiento-Márquez, E., Rodríguez-Triana, M. J., Wagner, M., & Ley, T. (**2024**). Mirror mirror on the wall, what is missing in my pedagogical goals? The impact of an AI-driven feedback system on the quality of teacher-created learning designs. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 145-156). New York, NY, USA: ACM.
63. Khaldi, A., Bouzidi, R., & Nader, F. (**2023**). Gamification of e-learning in higher education: A systematic literature review. *Smart Learning Environments*, *10*(1), 10-21.
64. Kopp, F., & Mendell, J. T. (**2018**). Functional classification and experimental dissection of long noncoding RNAs. *Cell*, *172*(3), 393-407.
65. Lv, B. Q., Weng, H. M., Fu, B. B., Wang, X. P., Miao, H., Ma, J., Richard, P., Huang, X. C., Zhao, L. X., Chen, G. F., Fang, Z., Dai, X., Qian, T., & Ding, H. (**2015**). Experimental discovery of Weyl semimetal TaAs. *Physical Review X*, *5*(3), 031013–031026.
66. Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (**2017**). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, *69*(12), 371-380.
67. Baghestani, A. R., Ahmadi, F., Tanha, A., & Meshkat, M. (**2019**). Bayesian critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, *52*(1), 69-73.
68. Alavi, M., Biros, E., & Cleary, M. (**2024**). Notes to factor analysis techniques for construct validity. *Canadian Journal of Nursing Research*, *56*(2), 164-170.
69. Gamarra, M., Dominguez, A., Velazquez, J., & Páez, H. (**2022**). A gamification strategy in engineering education: A case study on motivation and engagement. *Computer Applications in Engineering Education*, *30*(2), 472-482.
70. Sotos-Martínez, V. J., Ferriz-Valero, A., García-Martínez, S., & Tortosa-Martínez, J. (**2024**). The effects of gamification on the motivation and basic psychological needs of secondary school physical education students. *Physical Education and Sport Pedagogy*, *29*(2), 160-176.
71. Gkintoni, E., Vantaraki, F., Skoulidi, C., Anastassopoulos, P., & Vantarakis, A. (**2024**). Promoting physical and mental health among children and adolescents via gamification: A conceptual systematic review. *Behavioral Sciences*, *14*(2), 102-134.
72. Monroe, C. M., Zosel, K., Stansbury, M., Younginer, N., Davis, R. E., Dutton, G., Newton Jr., R. L., Cai, B., & West, D. S. (**2024**). A focus group study among insufficiently physically active African American adults regarding technology-delivered team-based gamification for physical activity promotion. *mHealth*, *10*(1), 3-3.
73. Torres-Toukoumidis, A., León, D. V., De-Santis, A., & López-López, P. C. (**2022**). Gamification in ecology-oriented mobile applications: Typologies and purposes. *Societies*, *12*(2), 42-53.
74. Saleem, A. N., Noori, N. M., & Ozdamli, F. (**2022**). Gamification applications in e-learning: A literature review. *Technology, Knowledge and Learning*, *27*(1), 139-159.
75. Flavián, C., Ibáñez-Sánchez, S., Orús, C., & Barta, S. (**2024**). The dark side of the metaverse: The role of gamification in event virtualization. *International Journal of Information Management*, *75*(23), 102726–102738.
76. Glikson, E., & Woolley, A. W. (**2020**). Human trust in artificial intelligence: A review of empirical research. *Academy of Management Annals*, *14*(2), 627-660.
77. Chen, L., Chen, P., & Lin, Z. (**2020**). Artificial intelligence in education: A review. *IEEE Access*, *8*(1), 75264–75278.
78. Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (**2021**). Artificial intelligence in drug discovery and development. *Drug Discovery Today*, *26*(1), 80-93.
79. Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., & Shen, D. (**2021**). Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. *IEEE Reviews in Biomedical Engineering*, *14*(1), 4-15.
80. Zhang, C., & Lu, Y. (**2021**). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*, *23*(3), 100224-100238.
81. Okunaiya, O., Austin, R., & Zhu, S. Y. (**2024**). ChatGPT-enabled network automation using API-based prompts. In *NOMS 2024-2024 IEEE Network Operations and Management Symposium* (pp. 1-5). Seoul: IEEE.
82. Lappalainen, Y., & Narayanan, N. (**2023**). Aisha: A custom AI library chatbot using the ChatGPT API. *Journal of Web Librarianship*, *17*(3), 37-58.
83. Chen, B. H., & Chen, C. C. (**2023**). Invention of Line-ChatBot: An innovative application of ChatGPT API and LINE Bot for enhanced student learning. In *2023 IEEE 6th International Conference on Knowledge Innovation and Invention (ICKII)* (pp. 452-455). Sapporo: IEEE.
84. Abasi, A., Hoseinabadi, R., Raji, P., Friedman, J. H., & Hadian, M. R. (**2022**). Evaluating oculomotor tests before and after vestibular rehabilitation in patients with Parkinson's disease: A pilot pre-post study. *Parkinson's Disease*, *2022*(1), 1-6.
85. Craig, S., Stark, P., Wilson, C. B., Carter, G., Clarke, S., & Mitchell, G. (**2023**). Evaluation of a dementia awareness game for undergraduate nursing students in Northern Ireland: A pre-/post-test study. *BMC Nursing*, *22*(1), 177-1.

86. Al-Mughairi, H., & Bhaskar, P. (**2024**). Exploring the factors affecting the adoption of AI techniques in higher education: Insights from teachers' perspectives on ChatGPT. *Journal of Research in Innovative Teaching & Learning*, *17*(1), 110-123.
87. Li, T., Ji, Y., & Zhan, Z. (**2024**). Expert or machine? Comparing the effects of pairing student teachers with in-service teachers and ChatGPT on their critical thinking, learning performance, and cognitive load in an integrated-STEM course. *Asia Pacific Journal of Education*, *44*(1), 1-16.
88. Guo, Y., & Lee, D. (**2023**). Leveraging ChatGPT for enhancing critical thinking skills. *Journal of Chemical Education*, *100*(12), 4876–4883.
89. Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (**2023**). Can large language models provide feedback to students? A case study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 323-325). Orem, UT, USA: IEEE.
90. Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (**2024**). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*(1), 1-15.
91. Abolnejadian, M., Alipour, S., & Taeb, K. (**2024**). Leveraging ChatGPT for adaptive learning through personalized prompt-based instruction: A CS1 education case study. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-8). New York, NY, USA: ACM.
92. Rusandi, M. A., Ahman, Saripah, I., Khairun, D. Y., & Mutmainnah. (**2023**). No worries with ChatGPT: Building bridges between artificial intelligence and education with critical thinking soft skills. *Journal of Public Health (Bangkok)*, *45*(3), e602–e603.