Research Article

© 2026 Rojas Hilario et al.
This is an open access article licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (https://creativecommons.org/licenses/by-nc/4.0/)

Received: 15 February 2026 / Accepted: 24 August 2026 / Published: September 2026

The Intervention Programme Based on Chatbots "Brain in Action" and the Development of Critical Thinking among Secondary School Students in Lima, 2025

Exalto Celso Rojas Hilario

Wilfredo Carcausto-Calla

César Vallejo University, Lima, Peru

DOI: https://doi.org/10.36941/jesr-2026-0414

Abstract

This study examines the impact of the chatbot-based intervention programme Brain in Action (PIBIAC) on the development of critical thinking among secondary school students in Lima, Peru, aligned with Sustainable Development Goal 4 (SDG 4). Adopting a quantitative approach and a quasi-experimental design with control and experimental groups (n=62), the study evaluated the effectiveness of chatbots as mechanisms of cognitive mediation. Data were collected using an instrument that measures critical thinking across five dimensions and analysed through descriptive and inferential statistics, specifically the Mann–Whitney U test. The findings indicate that the implementation of PIBIAC produced a statistically significant effect. A significant improvement (P<0.001) was observed compared to the control group, with better logical-deductive performance, inferential capacity, and evaluation of arguments. The study concludes that the chatbot-based intervention programme "Brain in Action" had a significant influence on the critical thinking of secondary school students at an educational institution in Lima during 2025.

Keywords: *Chatbot, critical thinking, artificial intelligence, secondary education, critical sense, cognition.*

1. Introduction

Developing critical thinking (CT) involves using advanced intellectual abilities that allow one to support ideas with relevant and well-founded information. As Castillo Cuadra (2020) points out, this skill is essential for young people to have tools that help them to better analyse and understand the reality that surrounds them. Globally, this competence has become the most relevant for secondary school students in the 21st century, as it enables them to face various challenges in different areas of their lives. Along the same lines, the United Nations Educational, Scientific and Cultural Organisation (UNESCO, 2023) highlighted the urgency of incorporating teaching approaches that cut across all areas of knowledge, in order to prepare adolescents to address current problems from a reflective and well-founded population.

However, there remains a marked separation between what is proposed in educational discourse and what actually happens inside classrooms. The Inter-American Development Bank

(IDB, 2018) pointed out that, although critical thinking is usually explicitly included in curricula, its effective application in everyday teaching practices continues to be limited. This situation is supported by concrete figures: only 30% of secondary school students manage to demonstrate an adequate level of critical judgement, which highlights a profound crisis in learning that affects different countries, regardless of their borders, Organisation for Economic Co-operation and Development (OECD, 2023).

Nevertheless, a profound disconnect persists between pedagogical discourse and classroom practice. The IDB (2018) warned that, although critical thinking (CT) is often explicitly integrated into curricula, its substantive implementation in teaching dynamics remains scarce. This deficiency is reinforced by empirical data revealing that barely 30% of secondary students reach a competent level of discernment; this highlights a learning crisis that transcends national borders (OECD, 2023).

Currently, educational systems continue to place greater importance on the repetitive memorisation of information than on the in-depth analysis of ideas, which represents an obstacle to the strengthening of PC. UNESCO (2024) has pointed out a notable lag in the updating of teaching content: although practically all countries recognise which are the fundamental competencies for the future, only 11 out of 51 have managed to incorporate the use of Artificial Intelligence into their curricular proposals. Added to this situation are inequalities in access to the internet and digital tools, factors that make it difficult to move towards a model of education adapted to each student, especially in those environments with fewer economic and social resources.

The lack of these skills directly affects the future opportunities of new generations. According to the United Nations Children's Fund (UNICEF, 2021), the absence of critical thinking exposes young people to situations of vulnerability, as it limits their ability to face risks, integrate into the world of work, and solve everyday problems. For its part, Sustainable Development Goal 4 (SDG4) urgently emphasises the importance of strengthening this form of reasoning, so that students can question existing structures and defend well-founded positions within their community. Although many countries have redesigned their curricula by adopting a competency-based approach, the real and profound implementation of these models continues to be a significant challenge.

In the Latin American context, this situation becomes even more critical. Various organisations have indicated that educational systems should offer high-quality services that reduce the existing gaps between different regions. Nevertheless, the effective incorporation of PC faces considerable obstacles, mainly due to the lack of curricula that explicitly include it and insufficient teacher training in this area, according to the Economic Commission for Latin America and the Caribbean (ECLAC, 2023). Added to this is the fact that, globally, almost half of teachers do not use tools such as artificial intelligence to promote the ability to argue among their students. As a result, a vertical teaching model is perpetuated, in which the pupil remains a passive recipient of information (UNESCO, 2024).

In Peru, educational data reveal an alarming situation: only 18% of students manage to analyse and question the information they receive in a well-founded manner (Ministry of Education, MINEDU, 2023), while reasoning skills reach barely 1.7% and active participation in classes only 2.0% (MINEDU, 2022). These figures warn of the risk of forming a citizenry with limited critical capacity. According to the National Education Council (CNE, 2022), the origin of this problem lies in a teaching model based on the passive transmission of knowledge, where the pupil is limited to listening, copying and memorising, without developing processes of analysis or critical reflection.

Although the National Basic Education Curriculum of Peru (CNEB, 2016) established the development of PC and creativity as a priority to prepare students for the demands of the labour market in 2025, the results obtained in the PISA 2022 test reveal a very different reality. With an

average of just 23.5 points and more than half placed at a basic level of performance, this study confirms that the curriculum is disconnected from what actually happens in the classrooms. There, teaching practices remain traditional, and deep equity gaps persist, affecting learning (Vivanco Enríquez, 2026).

This issue becomes apparent in regions such as Puno, where local research reveals that only 13% of students achieve an adequate performance in CT (Atencio, 2021). In the specific case of Lima, studies such as that by Saavedra-Pizarro (2024) confirm insufficient levels of this ability, with notable difficulties in interpreting and evaluating arguments. This situation ultimately promotes passive, rote-based learning, which makes young people more vulnerable to misinformation and limits their autonomy to build successful life projects.

This situation is evident in critical thinking among secondary school students in the district of San Juan de Lurigancho, who show serious difficulties in reasoning logically, constructing solid arguments, and making decisions for themselves. This limitation encourages a passive and rote-learning attitude towards learning, which makes them more vulnerable to misinformation and hinders the construction of their own life projects. Faced with this worrying outlook, it is urgent to design and implement an intervention programme based on the use of conversational artificial intelligence, such as chatbots, in order to systematically strengthen young people's cognitive skills and intellectual autonomy.

The difficulty students show in supporting their ideas with concrete evidence and making well-founded decisions demonstrates serious shortcomings in teaching methods, which weakens their intellectual curiosity and makes them more prone to fall into misleading reasoning or fallacies. This stagnation in the development of autonomous learning limits access to quality education, which highlights the urgent need to carry out quantitative research with follow-up over time. In response to this situation, the “Brain in Action” Intervention Programme with chatbots (PIBIAC) is proposed, designed to strengthen in a structured and progressive manner the ability to argue, reflect and critically evaluate among secondary school students. Faced with this gap in knowledge, the following research question arises:

In what way did PIBIAC influence CT in secondary school students of an educational institution in Lima during the year 2025? Likewise, at a specific level: In what way did PIBIAC influence the development of CT in inference, evaluation of arguments, deduction, supposition, and interpretation in secondary school students of an educational institution in Lima during the year 2025?

The theoretical justification for this study is based on its aim to reduce the deficiency in PC training through the PIBIAC programme, which proposes a learning model supported by conversational artificial intelligence. Inspired by the scaffolding theory, this model enables students to move from a passive receptive attitude to genuine cognitive autonomy, thanks to immediate feedback tailored to each individual. By combining pedagogical approaches with intelligent technology, this research addresses a significant gap in current knowledge, providing concrete data on the usefulness of chatbots for facilitating complex thinking processes and thus contributing to the international debate on the incorporation of digital tools into the secondary education curriculum (Laun and Wolff, 2025).

The methodological justification for this research lies in the use of a quasi-experimental design accompanied by structured instruments, which allows for precise measurement of the effect of the intervention programme in real learning situations. This approach not only ensures the collection of objective information, but, as Tobón (2027) points out, also encourages students to engage in processes of reflection and awareness regarding their own thinking. Additionally, the study enriches academic knowledge by demonstrating how quantitative methodologies are integrated with emerging technologies such as artificial intelligence, providing concrete evidence in a research area that is still in its early stages (Kurmanbayeva and Besekeeva, 2025).

The practical justification for this research is expressed in the application of the PIBIAC

programme, a technological tool that changes the classroom dynamic by assigning the chatbot the task of offering immediate and personalised feedback to each student. This enables teachers to overcome the limitations of the traditional teaching model, make better use of class time, and reduce the passive attitude that characterises students in contexts such as San Juan de Lurigancho. In summary, this intervention has succeeded in shifting from an approach based on repetitive memorisation to a model that prioritises student autonomy and the ability to solve complex problems (Tobón, 2017; Kurmanbayeva and Besekeeva, 2025).

From an epistemological perspective, this research is based on Karl Popper's critical rationalism, a school of thought which maintains that knowledge advances when ideas or hypotheses are put to the test and are refuted when they do not fit the facts. Following this approach, the study evaluates the effectiveness of the PIBIAC programme in a practical and measurable way, thus reproducing the same process of questioning, analysing, and contradicting arguments that it seeks to develop in students. In this way, by promoting CR among young people, a rational attitude is encouraged, aimed at constantly improving knowledge through the use of rigorous logic and continual comparison with reality (Humpiri et al., 2021).

Formally, the overall formal objective was: To determine the influence of the PIBIAC programme on the development of Critical Thinking among secondary school students in an educational institution in Lima during the year 2025. In addition, the specific objectives were established as follows: To determine the influence of PIBIAC on the development of critical thinking in inference, evaluation of arguments, deduction, supposition, and interpretation, among secondary school students in an educational institution in Lima during the year 2025.

Various international studies have shown that PC is not a fixed ability, but can be developed and change according to environmental conditions. For example, Köksal and Çöğmen (2018) found that skills such as evaluating information and making inferences vary according to characteristics unique to each individual. Likewise, Mendoza et al. (2016) demonstrated, through an experiment, that it is possible to achieve significant improvements: in general terms, student groups increased their performance by up to 29.1%, while at an individual level some people achieved progress of 66.6% in abilities such as explaining ideas and obtaining logical conclusions. These results confirm that PC is a malleable skill, whose strengthening depends directly on the use of educational strategies specifically designed to enhance the most complex mental processes.

The use of information technologies and artificial intelligence has ceased to be simply an element that motivates students and has become a true driver of thinking, as it allows for the strengthening of autonomy and the mastery of complex skills through the provision of immediate feedback (Sperling et al., 2024). Previous research, such as that by Medina Salazar (2017) and Romaní et al. (2021), has shown that well-structured programmes generate significant progress in general competences, enabling up to 36.3% of students to reach higher levels of performance. Therefore, the application of planned interventions not only develops deep cognitive abilities, but also reinforces the student's intellectual confidence and the positive perception they have of their own learning.

According to Dávila Newman (2006), deductive reasoning is fundamental for constructing logical arguments which, starting from general principles and scientific laws, allow valid conclusions to be reached. In this vein, Paz et al. (2010) found statistically significant improvements in induction and deduction abilities after applying certain educational interventions ($p=0.003$). However, other skills such as observation (0.0547) and the formulation of assumptions (0.687) did not show significant changes. These results demonstrate that the different dimensions of critical thinking do not develop at the same pace nor respond with equal intensity to teaching strategies, which highlights the diverse and complex nature of this type of learning.

According to De la Cruz et al. (2023), the comprehension of written texts requires processes

of analysis and interpretation that are fundamental for handling complex information. However, Padilla-Partida (2023) warned that inappropriate use of tools such as ChatGPT can reduce students' ability to summarise ideas and organise their thinking. This close relationship between reading comprehension and artificial intelligence indicates that, when technology is used as a substitute for personal analysis rather than as a support, there is a risk of weakening the foundations of RC. For this reason, the PIBIAC programme is presented as a cognitive stimulation tool designed to strengthen human reasoning, not to replace it.

Castillo Herrera (2023) found that the majority of students have a positive view of generative artificial intelligence, as they consider it a tool that facilitates PC and problem-solving: 56% of young people indicated that this contribution occurs "frequently", while 8% rated it as "constant". These results show that, although there is a danger that students may become dependent on technology, there is also a favourable disposition towards its use. This finding highlights the importance of the PIBIAC programme, as it provides the appropriate framework to guide these positive perceptions towards an organised and truly beneficial pedagogical application.

National studies reveal a critical perspective regarding the development of Critical Thinking (CT) in schools. Saavedra-Pizarro (2024) reported deficient levels in Lima, highlighting significant gaps in interpretation (45.7%) and the evaluation of arguments (43.7%); these figures coincide with the findings of Álvarez et al. (2020), where 80% of students are in the initial stages of reasoning. Furthermore, Castillo (2023) confirms the correlation between the identification of assumptions and reading comprehension, warning that poor analysis of premises seriously compromises both textual decoding and reflection.

Unlike the deficiencies observed in the country's capital, the results obtained in regions such as Tacna demonstrate the great value that technological solutions can have when they are applied appropriately. Ledesma and Sevairos (2023) found that the development of PC is closely linked to the planned and strategic use of information technologies, and not simply to their occasional use or lack of method. This precedent confirms that a well-structured intervention, such as the PIBIAC programme, has the capacity to generate significant pedagogical progress and foster high-level cognitive skills, thus managing to overcome the low levels of competence that currently prevail in much of the national education system.

These results are very similar to those obtained by Espinal et al. (2022) in their own research. In that study, the authors managed to reduce the percentage of students who were at a low level of argument evaluation, which dropped from 70% to 23.3%; at the same time, the group of pupils at a developing level rose to 53.3%. Something similar occurred in the capacity for interpretation, where the initial low level was drastically reduced from 56.7% to 1%, which allowed 13% of the students to reach a developing level and 15% to achieve an accomplished level. These data confirm that critical thinking is a skill that can be modified and improved, and that it responds favourably to specific teaching strategies that are well planned and systematically applied.

Research such as that by Chacón et al. (2023) and Castro et al. (2023) supports these findings by demonstrating that well-planned educational strategies, such as Project-Based Learning (PBL), produce positive and statistically significant effects on the development of critical thinking. In concrete terms, these interventions resulted in the percentage of students with an achieved level of critical thinking rising from 6.7% to 26.7%, while those at a low level were drastically reduced from 73.3% to 23.3%. This evidence indicates that careful pedagogical design, whether project-based or supported by technological tools, ensures sustained improvements in students' critical capacity, provided it is applied in organised classroom environments with a clear methodology.

Many theories have sought to explain what reflective practice entails. One of the most influential is the theory of reflective thinking proposed by Dewey (1933), which remains relevant today, although it has been reinterpreted over time. Thus, what was originally understood as an individual 'Reflective Practice' (2010) has more recently given way to a collaborative 'Civic and

Digital Inquiry' (2020). However, despite this shift in emphasis, the cyclical mechanism proposed by Dewey remains intact: the process begins with a situation of uncertainty or doubt; the problem is then clearly identified and defined; possible hypotheses or solutions are then proposed; the consequences of each are then considered; and finally, they are tested through action and verification with concrete facts.

The theory of critical thinking proposed by Halpern (1988) defines it as a set of mental abilities and attitudes that must be taught explicitly, through deliberate practice and in a variety of learning contexts. Today, this view has evolved and places a strong emphasis on metacognitive monitoring, that is, the student's ability to observe and regulate their own thought process. This skill is now considered essential for identifying and correcting the biases or errors in reasoning that arise when navigating digital environments full of complex information, thereby enabling conscious and self-regulated decision-making.

Furthermore, the theory of critical thinking developed by Ennis (1987) defines it as a form of sensible and reflective reasoning, aimed at making informed decisions about what to believe or what to do; this approach incorporates both analytical skills and personal attitudes, such as a willingness to consider different points of view. Over time, this theory evolved and, by 2028, had undergone a significant shift towards the digital sphere. Today, assessing whether a source is reliable is no longer limited to traditional media, but focuses on addressing phenomena such as the mass dissemination of false information (viral disinformation) and the effects of algorithms that can influence what we see and think.

Similarly, Facione's theory of critical thinking, which defines it as a self-regulated judgement (1990; 2020), maintains that critical thinking is a voluntary and reflective human action that is particularly activated when we find ourselves in uncertain or ambiguous situations. This model functions thanks to the inseparable integration of two fundamental components: on the one hand, the mastery of six cognitive skills, namely interpretation, analysis, evaluation, inference, explanation and self-regulation; and, on the other hand, the activation of effective dispositions or intellectual virtues, such as curiosity, honesty and open-mindedness.

Similarly, the theory of critical thinking developed by Paul and Eder (2003) is based on an initial proposal that sought to describe the structure of reasoning through so-called 'elements of thought' and to establish quality criteria through universal intellectual standards, whilst also distinguishing between 'weak' and 'strong' critical thinking. The current version of this theory retains this structure, but differs from the previous model in that it places greater emphasis on the cultivation of intellectual virtues, such as the humility to recognise one's own limitations and the autonomy to think for oneself, as well as on the application of this theoretical framework to ethical dilemmas and the ways in which information can be manipulated in the civic and digital spheres.

Furthermore, Watson and Glaser's (2012) psychometric theory is based on the measurement of logical reasoning across five main dimensions: the ability to make inferences, the recognition of implicit assumptions, deductive reasoning, the interpretation of information, and the evaluation of arguments. In its contemporary form, this theory has evolved and no longer focuses solely on academic assessment, but rather on predictive validation in professional settings. As a result, the model has established itself as a key tool for linking the level of logical reasoning with job performance and efficiency within organisations.

In contrast to the above, various theories and approaches have examined the independent variable relating to chatbot-based intervention programmes. The evolution of these conversational assistants for educational purposes has accelerated significantly over the last ten years: building on the intelligent systems developed in the 1960s and the online tutoring that emerged in the 1990s, we now have complex tools integrated into learning platforms tailored to each student (Hobert, 2023). This technological advancement has been accompanied by pedagogical methodologies such as Project-Based Learning (PBL), collaborative learning and

human-computer interaction, all of which are fundamental to AI-supported teaching (Yim & Su, 2025). In short, this transformation has been driven by the need to adapt the digital environment to pedagogical models that require more agile, participatory and effective interaction (Moreira-Choez, 2025).

The theory of Connectivism, formulated by Siemens (2005) and Downes (2008), posits that learning essentially consists of forming networks and establishing connections between different points of information, including non-human sources such as digital databases. It operates on the ability to manage and nurture these connections in constantly changing environments, prioritising the capacity to find and link knowledge rather than accumulating it individually in memory. From this perspective, knowledge is understood as something distributed across the digital network, and this theory thus becomes a fundamental framework for technology-enhanced education and for managing the uncertainty inherent in the Internet age.

Similarly, the Theory of Cognitive Scaffolding (Wood et al., 1976), which is based on Vygotsky's (1978) concept of the Zone of Proximal Development, involves providing temporary support tailored to each student's needs, so that they can carry out tasks they would not be able to complete on their own. In recent years, this theory has evolved in two main directions: on the one hand, so-called 'distributed scaffolding' (Belland et al., 2024), and on the other, self-regulated learning (Acevedo, 2022). In these current approaches, artificial intelligence and intelligent programmes act as mediators, providing personalised support in real time. In this way, both content comprehension and complex metacognitive processes are reinforced, all within educational environments that integrate neuroeducation and digital tools.

In the same vein, the Mental Model Theory, proposed by Johnson-Laird (1983), argues that people do not reason primarily by applying abstract logical rules, but by constructing internal representations of meaning that simulate possible scenarios of reality. This theory has evolved thanks to contributions from researchers such as Ocampo Flórez (2027), who highlight the predictive capacity of these mental models to foster truly deep learning. Today, this approach allows for a diversification of ways to represent information within the classroom, making the theory a highly valuable teaching tool for students to understand complex concepts through the mental simulation of concrete situations.

Similarly, Schema Theory, proposed by Rumelhart (1980), posits that knowledge is stored in memory through mental structures or templates that group related information, enabling us to interpret new situations based on previous experiences. Initially, this theory was applied to the field of artificial intelligence and the comprehension of written texts, but over time it has evolved towards the teaching of mathematics through schema recognition (Jitendra and Lein, 2020). This update shows that, rather than mechanically memorising formulas, teaching students to identify patterns or underlying structures in complex problems helps to optimise their working memory and significantly improves their academic performance, particularly among those groups of pupils with learning difficulties.

Finally, the Theory of Situated Cognition, proposed by Brown et al. (1989), demonstrated that learning is a dynamic, social process that is deeply rooted in context and cannot be separated from the practice or the cultural environment in which it takes place. Unlike theories that focus on internal mental structures, such as schemas, this school of thought maintains that knowledge is constructed primarily through interaction with the environment. Its development, led by authors such as Durning et al (2020), has validated this approach in fields such as virtual simulations and vocational training, demonstrating that a person's experience or expertise does not depend solely on having abstract schemas in their mind, but also on their ability to interact with specific factors in the environment and on the quality of the physical and social environment in which they operate.

Furthermore, the general hypothesis of this study was as follows: The 'Brain in Action' chatbot-based intervention programme had a significant impact on the critical thinking skills of

secondary school students at an educational institution in Lima in 2025. Furthermore, at a specific level, the hypotheses were: The "Brain in Action" chatbot-based intervention programme had a significant influence on the development of critical thinking in the dimensions of inference, evaluation, deduction, supposition and interpretation among secondary school students at an educational institution in Lima during 2025.

2. Methodology

This research falls within the field of applied research and employs a quantitative approach, with the aim of improving educational practice by measuring the effect of the chatbot-based intervention programme (PIBIAC) on the critical thinking (CT) skills of secondary school students in Lima. To this end, a quasi-experimental design with equivalent groups was employed, which allowed for rigorous control of the variables involved and ensured that the results can be replicated in similar contexts (Hernández-Sampieri and Mendoza, 2023; Osmanović and Maksimović, 2022).

The research design was structured around two central aspects. On the one hand, the independent variable, the PIBIAC programme, consisted of a system that simulates educational dialogues in natural language and was implemented over ten face-to-face sessions. On the other hand, the dependent variable, namely the PC, was measured based on the model proposed by Watson and Glaser (2009), which considers several dimensions: inference, evaluation of arguments, deduction, identification of assumptions and interpretation.

The study population consisted of 169 fifth-year secondary school pupils in San Juan de Lurigancho, from whom a purposive sample of 62 pupils was selected. Furthermore, regarding the inclusion and exclusion criteria applied: For inclusion: These were fifth-year secondary school students, registered in the Educational Institution Management Support Information System (SIAGIE) for 2025, with consistent attendance. In addition, students had their guardians' permission to participate in the study. Exclusion criteria: Students from other educational levels did not participate in this study, nor did those removed by headteacher's decision, nor those not belonging to the educational institution in question, nor those not enrolled in SIAGIE 2025.

The sample was selected intentionally and defined as a representative segment drawn from the study population. It is from this subset that the necessary data is collected, ensuring that the findings can be directly linked to the entire target group (Hernández-Sampieri & Mendoza, 2023). The sample will consist of 62 students, of whom the 5th B class, with 32 students, will form the experimental group, and the 5th C class, with 30 students, will form the control group.

With regard to the data collection method, the use of a questionnaire was proposed; as Reyes (2022) notes, questionnaires are a key research tool in social studies, and the most appropriate approach is to ask the research subjects directly. The survey for this research was based on a questionnaire comprising 22 items, using a structured test that assesses five essential dimensions: inferring, evaluating arguments, deducing, assuming and interpreting.

An adapted and validated critical thinking test was used, derived from the classic instrument by Watson and Glaser (1994), which was originally developed to measure critical thinking in various populations. This instrument was previously adapted in Turkey by Orhan and Ceviker (2022), and subsequently contextualised and validated for the Peruvian context by Bernal et al. (2025) among secondary school students. Furthermore, the instrument was validated by eight experts, achieving a content validity of 0.945 (Aiken's V), with high scores for adequacy (0.988), clarity (0.979), coherence (0.913) and relevance (0.909), demonstrating the relevance of its items. In terms of reliability, it obtained an overall KR-20 coefficient of 0.709, indicating good internal consistency. The values per component range from 0.708 to 0.734, confirming the stability of the instrument.

For the descriptive analysis, the data were coded and organised in Excel or SPSS. Measures

of central tendency, frequency calculations and graphical representations will be carried out. Once the data have been organised, they will be analysed using SPSS version 27. In this regard, descriptive techniques such as the median, interquartile range, minimum and maximum were applied, in addition to box plots to compare the distribution of groups. Furthermore, the Shapiro-Wilk test was used to verify normality. For inferential analysis, the Mann-Whitney U non-parametric statistical test was used to test hypotheses; according to Sánchez Turcios (2015), these tests were used to compare the medians of independent groups with non-normal data distributions.

The PIBIAC intervention programme was delivered over ten face-to-face sessions, each lasting 45 minutes, and was structured in three phases: introduction, main body and conclusion. Each session addressed a key aspect of critical thinking, such as identifying hidden intentions, the different types of reasoning (deductive, inductive and abductive), recognising biases and fallacies, constructing rebuttals, analysing syllogisms, managing uncertainty using risk matrices, drawing up scenario trees, using abduction as the best explanation, and the triple interpretation of texts. During the course, students worked in groups and interacted with the chatbot "Lucho AI", which acted as a cognitive support by questioning their inferences, requesting textual evidence, refuting weak arguments, offering alternative perspectives and verifying the logical soundness of their answers.

The teacher acted as a guide: motivating students, explaining concepts, modelling procedures, supervising the use of the chatbot, facilitating the sharing of results, and encouraging reflection on their own learning. The students produced progressive outputs (worksheets, reports, matrices, maps, trees, reports) which were validated by the chatbot until they reached the "Achieved" or "Outstanding" levels. The implementation followed a set sequence: activation of prior knowledge, cognitive conflict, conceptualisation, guided practice, interaction with the chatbot, product development, sharing of adjustments, metacognitive reflection and self-assessment. This detailed description makes it possible to replicate the study in similar educational contexts.

Finally, the study complied with all the ethical requirements for research involving minors. Institutional authorisation was obtained via a formal document, in which Educational Institution 0158 "Santa María" granted permission, through its Headteacher, for the data generated in the study. Furthermore, informed consent was sought and obtained from the students, in a document that clearly and simply explains the purpose of the research, the procedures to be followed, the estimated time required for participation (approximately 40 minutes), the absence of risks or harm, the freedom to withdraw at any time without any consequences, and the assurance that the benefits will primarily be for the institution and public health, not for individual participants.

Furthermore, the confidentiality and anonymity of the data were ensured by coding the responses with identification numbers, thereby removing any link to the students' personal identities. The principal investigator safeguarded the data and undertook to delete it after a specified period. The fundamental ethical principles of autonomy (voluntary participation), non-maleficence (absence of harm), beneficence (contribution to educational knowledge) and justice (fair treatment and strict confidentiality) were upheld. The consideration of various ethical aspects guaranteed the integrity and credibility of the research, ensuring that all processes were respectful and fair (Taquette & Borges, 2022).

The chatbot tool was designed and built specifically for this research, based on the Socratic method and the use of scaffolding questions and challenging questions that promote critical thinking. In each session of the PIBIAC programme, a themed chatbot linked to the specific critical thinking content being covered was implemented. For its development, Gemini artificial intelligence was used as an assistant in generating questions and answers designed to progressively challenge the students' cognitive skills. The chatbot operated under a conditional

logic: when the student answered correctly, they received a positive message ("You did it") and automatically moved on to the next question or level. Conversely, if the answer was incorrect or insufficient, the chatbot provided explanatory feedback, clues or scaffolding questions, and remained on the same question until the student achieved a valid answer, at which point they moved on.

This design ensured a personalised, gradual and sustained learning process, in which each student progressed at their own pace, consolidating their reasoning and logical skills through guided interaction with the conversational system.

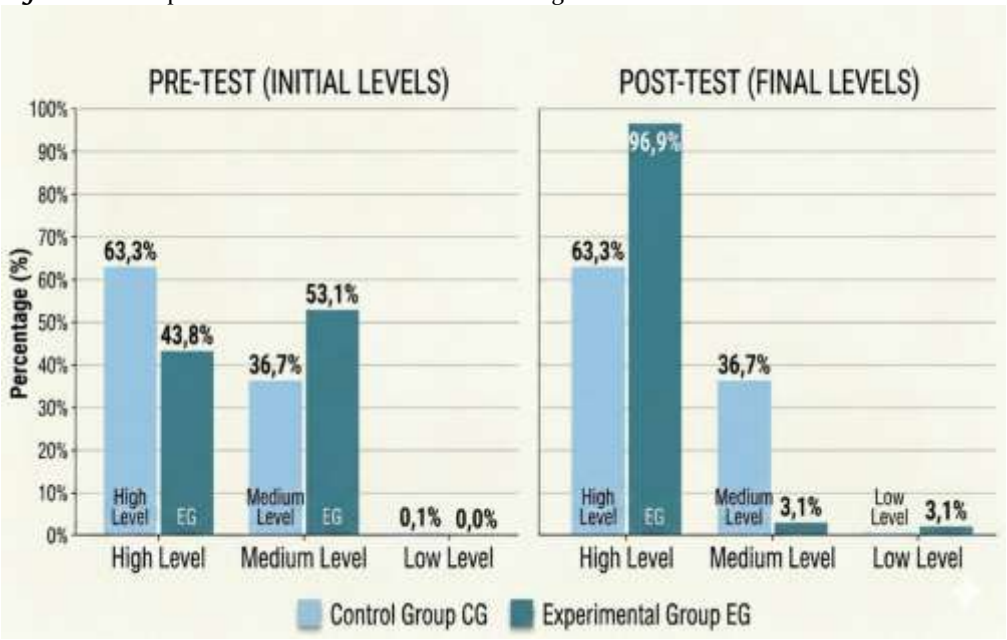
3. Results

The analysis presented the initial results for the CT variable, which were measured using a pre-test administered to an experimental group of 32 participants and a control group of 30 participants. The aim was to establish a baseline by identifying the distribution of participants across low, medium and high performance levels prior to any intervention. This initial assessment proved essential for measuring the subsequent impact of the educational programme.

3.1 Descriptive statistics

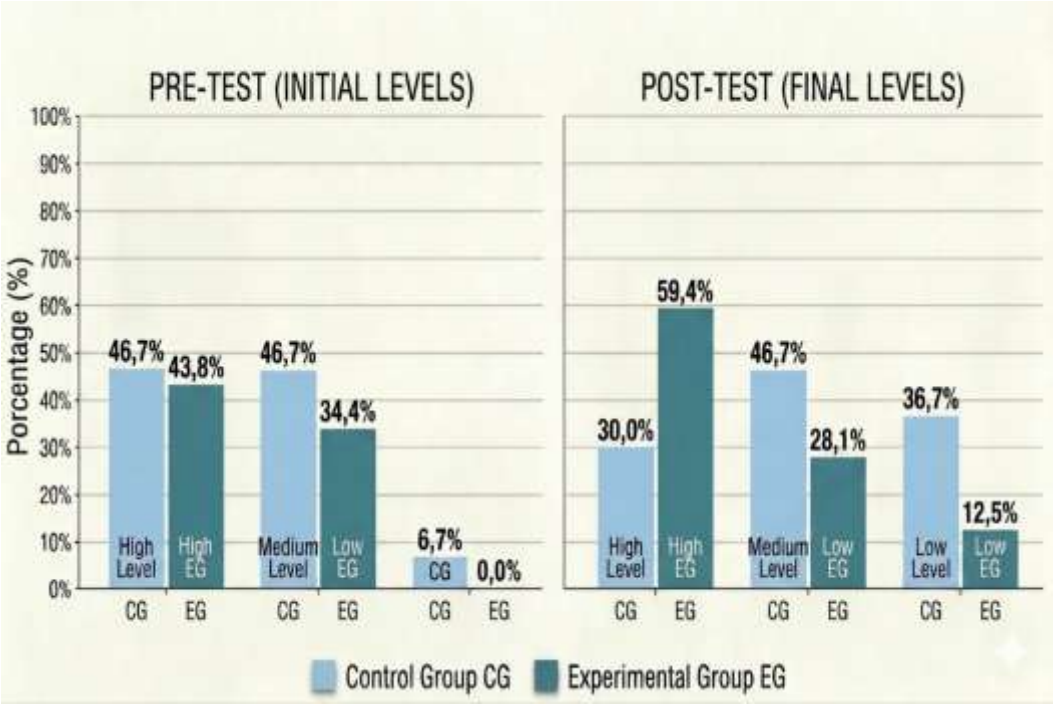
According to the data obtained for the CT variable in Figure 1, it was observed that, in the pre-test, the CG had a 63.3% advantage in the high and medium levels (36.7%) compared to the EG, which had 43.8% in the high level and 53.1% in the medium level. Following the intervention, the EG achieved a dramatic change in the post-test, with 96.9% of participants reaching the high level and 3.1% the medium level, massively outperforming the CG. The key finding was the complete stagnation of the CG, which maintained its initial distribution of 63.3% at the high level and 36.7% at the medium level in the post-test. This substantial difference and the consolidation of the EG irrefutably validated the programme's effectiveness in strengthening the CT.

Figure 1: Descriptive statistics for the critical thinking variable.



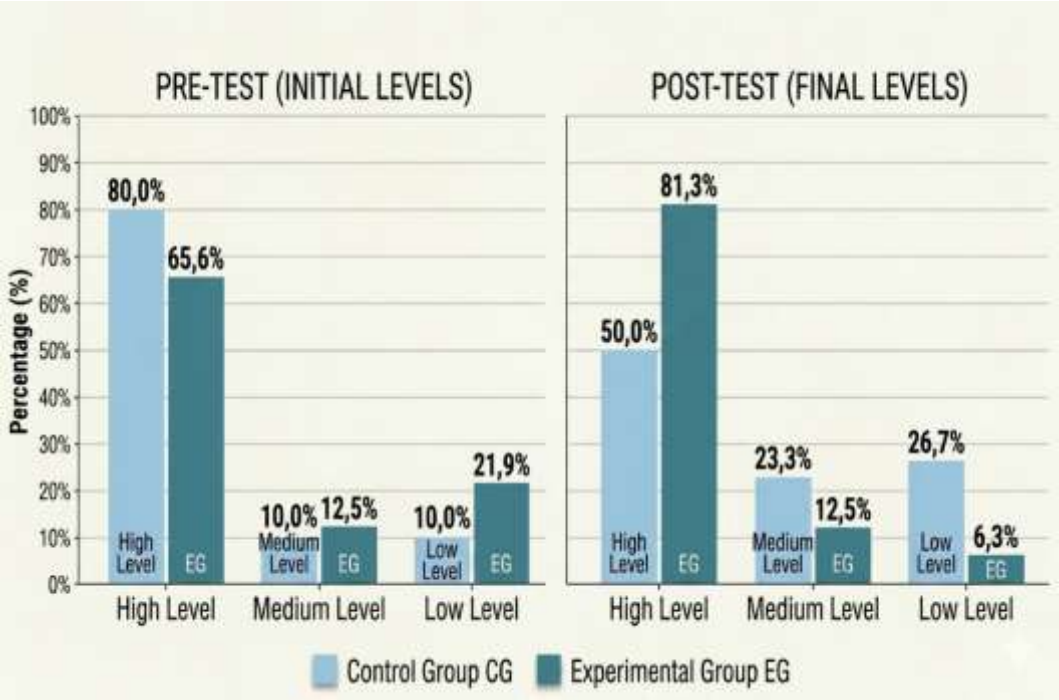
According to the data obtained on the inference dimension in Figure 2, the pre-test showed that the GC achieved a 46.7% advantage at the high and medium levels, and 6.7% at the low level, whereas in the GE, 43.8% of learners were at the high level, followed by 34.4% at the medium level. Similarly, in the post-test, the results showed positive progress in the GE, where 59.4% of students reached the high level, 28.1% the medium level and 12.5% the low level. However, the most significant finding was the regression in the GC, where the high level decreased by 16.7%. This deterioration correlated with a 30% increase in the low level, suggesting that, without intervention, inferential ability showed marked instability.

Figure 2: Descriptive statistics obtained from the inference dimension.



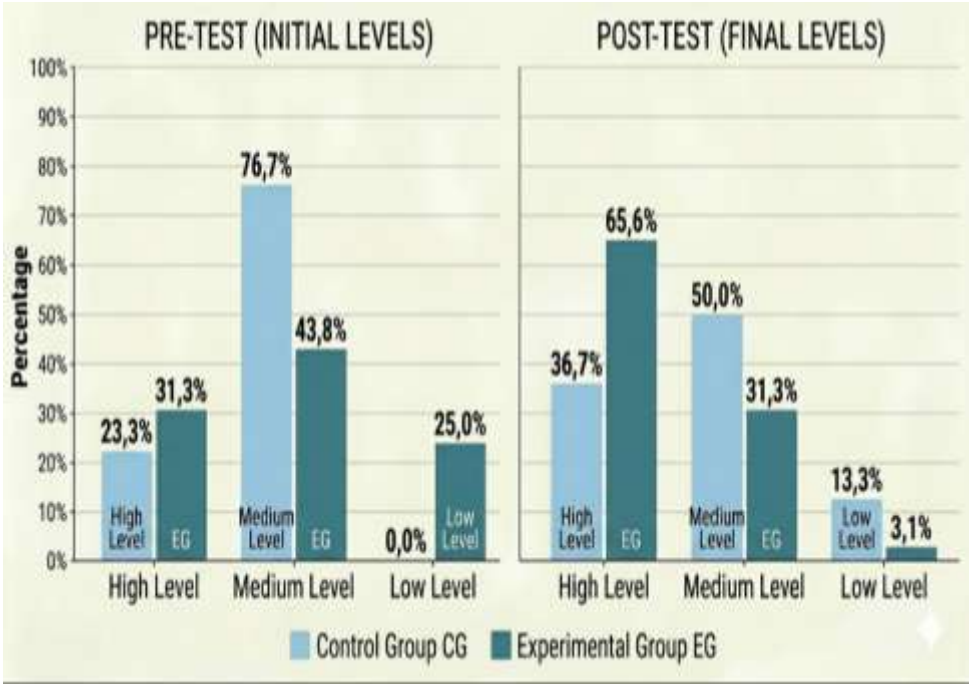
According to the data obtained on the 'argument evaluation' dimension in Figure 3, the pre-test revealed that, in the control group (CG), 80% of participants achieved the high level, with 10% at the medium and low levels; in contrast, in the experimental group (EG), 65.6% of students were at the high level, 12.5% at the medium level and 21.9% at the low level. Similarly, in the post-test, significant changes were evident in the EG, where 81.3% of students were at the high level, 12.5% at the medium level and only 6.3% at the low level, whereas in the CG, 50% of participants achieved the high level, 23.3% were at the medium level and 26.7% at the low level. The most significant finding was the sharp decline in the CG: its high level fell from 80% to 50% in the post-test, whilst the medium and low levels increased. This contrast demonstrated that the programme not only improved skills in the GE, but was also essential for consolidating and protecting competence, preventing the decline observed in the group without intervention.

Figure 3. Descriptive statistics obtained from the 'argument evaluation' dimension.



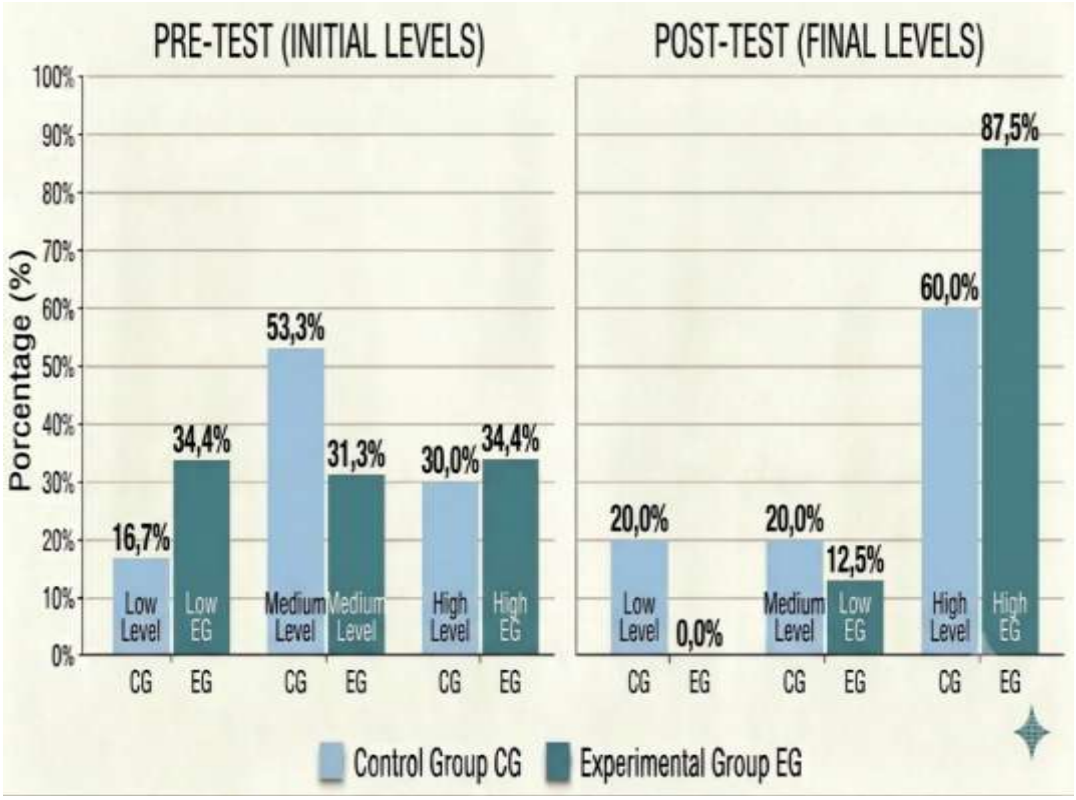
In Figure 4, for the 'inference' dimension, the pre-test results showed that the CG had 23.3% at the high level, 76.7% at the medium level and no participants at the low level, whereas the EG had 31.3% of participants at the high level, 43.8% at the medium level and 25.0% at the low level. In the post-test, the EG showed 65.6% of participants at the high level, 31.3% at the medium level and 3.1% at the low level, in contrast to the CG, which showed 36.7% at the high level, 50.0% at the medium level and 13.3% at the low level. Therefore, the extent of the improvement in the EG validated the programme's ability to develop deductive reasoning skills, even when starting from greater initial deficits.

Figure 4. Descriptive statistics obtained from the 'deduce' dimension.



In Figure 5 for the 'assumption' dimension, the pre-test results showed that the CG had 30% at the high level, followed by 53.3% at the medium level and 16.7% at the low level, whereas the EG had 34.4% of participants at the high level, 31.3% at the medium level and 34.4% at the low level, In the post-test, 87.5% of participants in the EG reached the high level, 12.5% remained at the medium level, and the low level was completely eliminated. In contrast, the CG showed a slight variation, with 60% at the high level and 20% at the medium and low levels. Therefore, the extent of the improvement in the EG confirmed that the programme acted as a powerful enhancer of competence.

Figura 5. Descriptive statistics obtained from the assumption dimension.



According to Figure 6 for the 'interpretation' dimension, the pre-test results showed that for GC, the results were slightly higher at the high level, with 46.7% of participants at the high level, 50.0% at the medium level and 3.3% at the low level, compared to 37.5% of participants at the high level, 46.9% at the medium level and 15.6% at the low level. In the post-test, a marked improvement was observed in the GE, where 65.5% of participants reached the high level, 28.1% remained at the medium level and a further 6.3% at the low level. In contrast, the GC showed a more moderate improvement, with 33.3% at the high level, 43.3% at the medium level and 23.3% at the low level. This divergence demonstrated that the intervention not only closed the initial gap but also completely reversed the proficiency hierarchy, proving the programme's ability to consolidate interpretation skills.

Figura 6. Descriptive statistics obtained from the 'interpret' dimension.

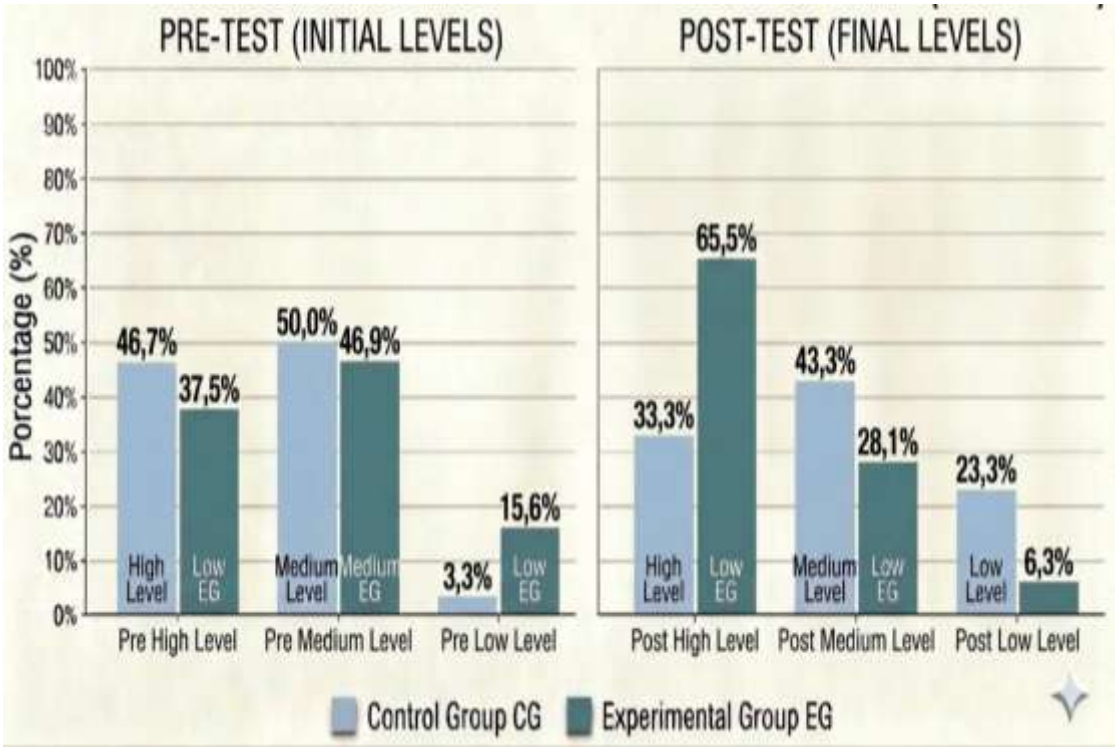


Table 1 shows that the Shapiro-Wilk test was used to verify the normality of the distributions: the 'critical thinking' variable in the experimental group followed a normal distribution, whereas in the control group it did not. Consequently, the non-parametric Mann-Whitney U test was used for the comparisons.

3.2 Hypothesis testing

3.2.1 Null hypothesis

Ho: The PIBIAC does not have a significant influence on the development of critical thinking among secondary school students at an educational institution in Lima in 2025.

Ha: The PIBIAC has a significant influence on the development of critical thinking among secondary school students at an educational institution in Lima in 2025.

3.3 Decision rule

Based on the two-tailed significance level
If $p < 0.05$, H_0 must be rejected.
If $p > 0.05$, H_1 must be accepted.
Based on the Z-score
If $Z \leq -1.96$ or $Z \geq 1.96$, the null hypothesis is rejected.

From this point onwards, the decision rule was applied to all tables. In addition, the control group was assigned the number 1 and the experimental group the number 2.

Table 1. Shapiro-Wilk Normality Test for the Critical Thinking Variable.

Variable	M	Estadístico	gl	Sig.
Critical thinking	Control	,923	30	,033
	Experimental	,934	32	,052

The findings presented in Table 2 for the pre-test showed a similar average range (34.52 and 28.67) in groups 1 and 2. Furthermore, the two-tailed asymptotic p-value for the pre-test was (0.200 > 0.05), indicating that the differences were not statistically significant. In contrast, in the post-test, a clear difference was observed in the mean ranges (21.92 and 40.48), with a two-tailed p-value of (0.000 < 0.05). This result demonstrated a significant difference between the GC and GE groups, allowing the null hypothesis to be rejected and the alternative hypothesis to be accepted. These results revealed that the implementation of the PIBIAC produced a significant improvement in the development of CT in secondary school students in Lima.

Table 2. Results of the Mann-Whitney U test for the null hypothesis.

	Ranks		
	Groups	N	Average range
	1	30	34,52
Pre-test	2	32	28,67
	Total	62	
	1	30	21,92
Post-test	2	32	40,48
	Total	62	
	Contrast statistics		
	Pre-test		Post-test
Mann-Whitney U test	389,500		192,500
Z	-1,280		-4,084
Sig. Asymptotic (Bilateral)	,200		,000

3.4 Grouping variable: Group

The findings presented in Table 3 for the pre-test showed a similar average range (33.30 and 29.81) in groups 1 and 2. Furthermore, the two-tailed asymptotic significance in the pre-test was 0.407 (> 0.05), meaning the differences were not significant. In contrast, in the post-test, a clear difference was observed in the average ranges (25.68 and 36.95) in the CG and EG, with a two-tailed significance of 0.008 (< 0.05). This result demonstrated a significant difference between the GC and GE groups, allowing the null hypothesis to be rejected and the alternative hypothesis to be accepted. These results demonstrate that the implementation of the PIBIAC produced a significant improvement in the development of critical thinking for inference among secondary school students in Lima.

Specific Hypothesis Test 1

Ho: The PIBIAC does not have a significant influence on the development of critical thinking

skills for inference among secondary school students at an educational institution in Lima in 2025.

Ha: The PIBIAC has a significant influence on the development of critical thinking skills for inference among secondary school students at an educational institution in Lima in 2025.

Table 3. Results of the Mann-Whitney U test for specific hypothesis 1.

	Ranks		
	Groups	N	Average range
	1	30	33,30
Pretest	2	32	29,81
	Total	62	
	1	30	25,68
Postest	2	32	36,95
	Total	62	
	Contrast statistics		
	Pretest		Postest
Mann-Whitney U test	426,000		305,500
Z	-,829		-2,635
Sig. Asymptotic (Bilateral)	,407		,008

4.5 Grouping variable: Group

The findings presented in Table 4 for the pre-test showed a similar average range (34.10 and 29.06) in groups 1 and 2. Furthermore, the two-tailed asymptotic significance in the pre-test was 0.161 (> 0.05), meaning the differences were not significant. In contrast, in the post-test, a clear difference was observed in the average ranges (26.30 and 36.38) in the CG and EG, with a two-tailed significance of 0.009 (< 0.05). This result demonstrated a significant difference between the GC and GE groups, allowing the null hypothesis to be rejected and the alternative hypothesis to be accepted. These results demonstrate that the implementation of the PIBIAC produced a significant improvement in the development of critical thinking for the evaluation of arguments among secondary school students in Lima.

Specific Hypothesis Test 2

Ho: The PIBIAC does not have a significant influence on the development of critical thinking skills for evaluating arguments among secondary school students at an educational institution in Lima in 2025.

Ha: The PIBIAC has a significant influence on the development of critical thinking skills for evaluating arguments among secondary school students at an educational institution in Lima in 2025.

Table 4. Results of the Mann-Whitney U test for specific hypothesis 2.

	Ranks		
	Groups	N	Average range
	1	30	34,10
Pretest	2	32	29,06

	Total	62	
	1	30	26,30
Postest	2	32	36,38
	Total	62	
	Contrast statistics		
	Pretest		Postest
Mann-Whitney U test	402,000		324,000
Z	-1,401		-2,620
Sig. Asymptotic (Bilateral)	,161		,009

4.6 Grouping variable: Group

The findings presented in Table 5 for the pre-test showed a similar average range (32.48 and 30.58) in groups 1 and 2. Furthermore, the two-tailed asymptotic significance in the pre-test was 0.665 (> 0.05), meaning the differences were not significant. In contrast, in the post-test, a clear difference was observed in the average ranges (25.42 and 37.20) in the GC and GE groups, with a two-tailed significance of 0.005 (< 0.05).

This result demonstrated a significant difference between the two groups, allowing the null hypothesis to be rejected and the alternative hypothesis to be accepted. These results demonstrate that the implementation of the PIBIAC produced a significant improvement in the development of critical thinking for deduction in secondary school students in Lima.

4.7 Specific Hypothesis Test 3

Ho: The PIBIAC does not significantly influence the development of critical thinking skills for deduction among secondary school students at an educational institution in Lima in 2025.

Ha: The PIBIAC significantly influences the development of critical thinking skills for deduction among secondary school students at an educational institution in Lima in 2025.

Table 5. Results of the Mann-Whitney U test for specific hypothesis 3.

	Ranks		
	Groups	N	Average range
	1	30	32,48
Pre-test	2	32	30,58
	Total	62	
	1	30	25,42
Post-test	2	32	37,20
	Total	62	
	Contrast statistics		
	Pre-test		Post-test
Mann-Whitney U test	450,500		297,500
Z	-,433		-2,817
Sig. Asymptotic (Bilateral)	,665		,005

4.8 Grouping variable: Group

The findings from the pre-test, presented in Table 4, showed a similar average range (31.93 and 31.09) in groups 1 and 2. Furthermore, the two-tailed asymptotic significance in the pre-test was 0.852 (> 0.05), meaning the differences were not significant. In contrast, in the post-test, a clear difference was observed in the mean ranks (25.70 and 36.94) in the CG and EG, with a two-tailed significance of 0.008 (< 0.05).

This result reflected a significant difference between the two groups, allowing the null hypothesis to be rejected and the alternative hypothesis to be accepted. These results demonstrate that the implementation of the PIBIAC produced a significant improvement in the development of critical thinking skills in secondary school students in Lima.

4.9 Specific Hypothesis Test 4

Ho: The PIBIAC does not significantly influence the development of critical thinking skills in secondary school students at an educational institution in Lima.

Ha: The PIBIAC significantly influences the development of critical thinking skills in secondary school students at an educational institution in Lima.

Table 4. Results of the Mann-Whitney U test for specific hypothesis 4.

	Ranks		
	Groups	N	Average range
	1	30	31,93
Pre-test	2	32	31,09
	Total	62	
	1	30	25,70
Post-test	2	32	36,94
	Total	62	
	Contrast statistics		
	Pre-test		Post-test
Mann-Whitney U test	467,000		306,000
Z	-,187		-2,645
Sig. Asymptotic (Bilateral)	,852		,008

4.10 Grouping variable: Group

The findings presented in Table 5 for the pre-test showed a similar average range (35.08 and 28.14) in groups 1 and 2. Furthermore, the two-tailed asymptotic significance in the pre-test was 0.120 (> 0.05), meaning the differences were not significant. In contrast, in the post-test, a clear difference was observed in the average ranges (25.67 and 36.97) in the GC and GE groups, with a two-tailed significance of 0.008 (< 0.05).

This result revealed a significant difference between the two groups, allowing the null hypothesis to be rejected and the alternative hypothesis to be accepted. These results demonstrate that the implementation of the PIBIAC produced a significant improvement in the development of critical thinking for interpretation among secondary school students in Lima.

4.11 Specific hypothesis test 5

Ho: The PIBIAC does not have a significant influence on the development of critical thinking skills for interpretation among secondary school students at an educational institution in Lima.

Ha: The PIBIAC has a significant influence on the development of critical thinking skills for interpretation among secondary school students at an educational institution in Lima.

Table 5. Results of the Mann-Whitney U test for specific hypothesis 5.

	Ranks		
	Groups	N	Average range
	1	30	35,08
Pretest	2	32	28,14
	Total	62	
	1	30	25,67
Post-test	2	32	36,97
	Total	62	
	Contrast statistics		
	Pre-test		Post-test
Mann-Whitney U test	372,500		305,000
Z	-1,553		-2,643
Sig. Asymptotic (Bilateral)	,120		,008

4.12 Grouping variable: Group

4. Discussion

With regard to the overall objective, the study empirically confirms that the PIBIAC programme has a significant impact on the critical thinking skills of secondary school pupils: the groups showed similar results in the pre-test (p=0.200), but the experimental group demonstrated a decisive improvement in the post-test (p=0.000), a result that replicates the advances reported by Köksal and Çöğmen (2018) and Saavedra (2024) with technological interventions. From a theoretical perspective, the chatbot's success is grounded in the Connectivism of Siemens and Downes (2005; 2008), which conceives of learning as the management of information networks that include non-human sources. On a practical level, a key caveat emerges: high effectiveness in a controlled environment may lead to technological dependency; therefore, it is recommended that the programme be complemented by teacher training that develops metacognitive autonomy outside the assisted environment.

In line with the first objective, regarding the inference dimension, the empirical findings indicate that the programme improved this ability: the pre-test showed no significant difference between groups (p=0.407), whereas the post-test revealed a significant difference (p=0.0008), consistent with Ocampo et al. (2025) and Castillo Herrera (2023). Theoretically, this improvement is explained by Cognitive Load Theory (Sweller, 1980; Pass and Kirschner, 2004), which posits that reducing irrelevant load frees up attentional resources for learning.

The key practical implication is that success may be due more to the simplicity of the algorithm than to a deep understanding of logical rigour; it is therefore suggested that the chatbot be combined with unassisted activities that require students to apply reasoning in less

structured contexts.

With regard to the second specific objective concerning the argument evaluation dimension, from an empirical perspective, the pre-test revealed no significant differences (p=0.161), whilst the post-test showed a significant improvement (p=0.009), in line with the findings of Saavedra (2024) and Espinal et al. (2022). The theoretical interpretation is based on Cognitive Scaffolding Theory (Wood, Bruner and Ross, 1976) and Vygotsky's Zone of Proximal Development (1978), updated by Belland et al. (2024) and Azevedo (2022) towards 'distributed scaffolding' using artificial intelligence. In practical terms, the risk is one of "incomplete transfer": the student learns to evaluate within the software, but may remain vulnerable to biases and misinformation when operating autonomously; therefore, teacher training must be geared towards internalising critical judgement outside the platform.

In accordance with the third specific objective regarding the deduction dimension, the empirical data reveal that the programme enhanced this skill: a significant post-test score (p=0.005) compared to the pre-test (p=0.665), in line with Mendoza et al. (2016) and Paz and Sánchez (2010). Theoretically, this is grounded in Johnson-Laird's (1983) Theory of Mental Models, which posits that deductive reasoning is based on internal semantic representations rather than formal rules. The practical implication is the danger of reducing deduction to an algorithmic process: students could become dependent on computational rigour and lack the flexibility to handle real-world ambiguities; it is recommended to include exercises with uncertain or incomplete premises.

According to the fourth specific objective regarding the inference dimension, the programme achieved a statistically significant improvement in the post-test (p=0.008) compared to the pre-test (p=0.852), as reported by Saavedra-Pizarro (2024) and Castillo et al. (2023). The theory underpinning this is Rumelhart's Schema Theory (1980): through guided questions, the chatbot activates and restructures prior cognitive structures to recognise implicit meanings. In practice, the risk is social decontextualisation: in reality, assumptions are linked to ideologies, values and emotions, whilst sterile training can result in a fragile ability to deal with emotionally charged or manipulative discourse; it is suggested that this be complemented with debates and analyses of real-life cases.

In accordance with the fifth specific objective regarding the interpretation dimension, the empirical findings show an improvement in the post-test (p=0.008) compared to the baseline (p=0.120), replicating the results documented by Saavedra-Pizarro (2024) and Espinal et al. (2022). Theoretically, this is grounded in the Theory of Situated Cognition (Brown, Collins and Duguid, 1989), which asserts that understanding arises from the interaction between the individual, the context and sociocultural practices. The practical implication is that the chatbot, by operating under a binary logic that minimises polysemy, may foster cognitive rigidity and a purely literal interpretation; it is advisable to complement the programme with tasks that require the handling of multiple meanings, contextual nuances and complex communicative intentions.

In terms of limitations and projections, from a methodological perspective, the small sample size and cross-sectional design limit the generalisability of the findings and our understanding of the durability of learning. Furthermore, the fact that the study was conducted in a single institution and that the chatbot was developed in-house without technical guidance restricts the ability to extrapolate the findings.

For future research, it is recommended to use longitudinal designs with large randomised samples across multiple institutions, to develop chatbots tailored to standardised technical parameters, and to include relevant measurements.

5. Conclusions

With regard to the overall objective, the empirical results confirm that the PIBIAC programme has a significant impact on the development of critical thinking among secondary school students in Lima 2025. The experimental group completely eliminated low-level performance, achieving 96.9% at the high level and 3.1% at the medium level, whilst the control group obtained an average of 21.92 compared to 40.48 for the experimental group (Z = -4.084; P = 0.000). These data validate the proposed hypothesis. However, it must be acknowledged that the study was limited to an educational intervention in Lima with a relatively small sample; therefore, these findings should not be generalised to other contexts without due caution.

With regard to the first specific objective concerning the inference dimension, the programme demonstrated a significant improvement empirically: 59.4% of the experimental group achieved a high level and 28.1% a medium level, with significantly different average scores between the groups (25.68 for the control group and 36.95 for the experimental group; Z = -2.635; P = 0.008). The hypothesis is thus confirmed, but it must be emphasised that these findings relate to the limited context of a school in Lima and a small sample, meaning that any extrapolation requires replication studies.

With regard to the second specific objective concerning the argument evaluation dimension, the empirical data also confirm the hypothesis: 81.3% of the experimental group scored at the high level and 12.5% at the medium level, with significant differences in mean scores (26.30 for the control group and 36.38 for the experimental group; Z = -2.620; P=0.009). The robustness of these results should not obscure the methodological limitation arising from the local nature of the study and the small sample size, which prevents the conclusions from being generalised without further studies in different populations and contexts.

With regard to the third specific objective—the deduction dimension—the programme had a significant impact: 65.6% of the experimental group achieved a high level, 31.3% a medium level and 3.1% a low level, with average scores of 25.42 for the control group and 37.20 for the experimental group; Z = –2.817; p = 0.005. The hypothesis is confirmed, albeit with the same caveat: the findings come from a single educational institution in Lima and a relatively small sample, so they should not be generalised without caution.

With regard to the fourth specific objective, the 'assumption' dimension, the results also support the hypothesis: 87.5% of the experimental group scored at the high level, 12.5% at the medium level, and the low level was eliminated, with significant differences in mean scores (27.70 for the control group and 36.94 for the experimental group; Z=-2.4645; p=0.008). Once again, it should be noted that these conclusions are limited to the specific context of the study (one institution, small sample). They should not be extrapolated without further research.

Finally, with regard to the fifth specific objective—the interpretation dimension—the programme showed a significant improvement: 65.5% of the experimental group achieved a high level, 28.1% a medium level and 6.3% a low level, with average scores of 25.67 for the control group and 36.97 for the experimental group; Z = –2.643; p = 0.008. The hypothesis is thus demonstrated, but the central limitation is reiterated: the study was conducted in a single educational institution in Lima with a relatively small sample, so the findings should not be generalised beyond this context without due caution. For future research, it is recommended to expand the sample, include multiple institutions and adopt longitudinal designs that allow for the evaluation of the sustainability of achievements and their transfer to diverse contexts.

6. Limitations and Future Directions

Despite the potential of the PIBIAC programme as a catalyst for critical thinking and its alignment with the objectives of SDG 4, this research identifies several limitations that will guide future work. The main limitation is that the study was confined to a single educational institution in Lima and involved a relatively small sample (n=62). Consequently, the findings should not be generalised without caution to other education systems with different structural, socio-economic or curricular realities. Furthermore, whilst the quasi-experimental design employed offers significant internal validity supported by the consistency of the effect sizes, it does not allow for complete control of the possible 'novelty effect' (or Hawthorne effect), that is, the initial improvement that could be attributed to students' enthusiasm for interacting with an artificial intelligence chatbot.

Consequently, future research should focus on longitudinal studies that enable the assessment of both the transfer of critical skills to unassisted contexts and the long-term retention of learning. Furthermore, it is essential to conduct comparative research between public and private educational settings, using larger and more randomised samples, in order to determine how the digital divide may influence the tool's effectiveness. Finally, it is suggested that the integration of large language models (LLMs) be explored, as these offer deeper and more personalised cognitive mediation, adapting to students' specific needs and reducing the risks of technological dependency or algorithmic rigidity highlighted in the discussion.

Statement on the use of artificial intelligence

During the preparation of this manuscript, the authors used Chat with DeepSeek AI as a support tool for (i) linguistic editing (improving clarity, cohesion and academic style), (ii) assistance in processing the results of statistical analysis (e.g., drafting concise explanations of diagnostic results and effect sizes based on the results generated by the researchers) and (iii) support for the interpretation of results by helping to structure alternative narrative frameworks consistent with the calculated findings. All statistical analyses were designed, performed and validated by the researchers, using the underlying datasets and standard statistical software, and no AI system performed calculations or generated numerical results. All citations and references were manually verified by the authors, including checking for accuracy and bibliographic relevance. The authors assume full responsibility for the content of the manuscript, the integrity of the analyses and the final interpretations presented.

References

Álvarez, M. Y., Menacho, I., Esquiagola, E. A., & Camarena, J. L. (2020*). Habilidades del pensamiento crítico en estudiantes de educación secundaria de Lima, Perú*. *INNOVA Research Journal,* 5(3.2), 97-110. https://doi.org/10.33890/innova.v5.n3.2.2020.1551

Atencio Chura, J. C. (2021). *Nivel de pensamiento crítico de los estudiantes de las Instituciones Educativas Secundarias "San Antonio de Padua" y "Carlos Rubina Burgos"-Puno*. https://repositorio.unap.edu.pe/items/06469e69-88f4-4424-85f2-48d6f8999162

Azevedo, R. (2022). *Lessons learned and future directions of MetaTutor: Leveraging multichannel data to scaffold self-regulated learning with an intelligent tutoring system.* Frontiers in Psychology, 13, Artículo 813632. https://doi.org/10.3389/fpsyg.2022.813632

Banco Interamericano de Desarrollo. (2018). *Los desafíos del pensamiento crítico en la educación del siglo XXI*. BID. https://blogs.iadb.org/educacion/es/educacionsigloxxi/

Bartley, J. E., Boeving, E. R., Riedel, M. C., Bottenhorn, K. L., Salo, T., Eickhoff, S. B., ... & Laird, A. R. (2018). *Meta-analytic evidence for a core problem solving network across multiple representational domains.* Neuroscience & biobehavioral reviews, 92, 318-337. https://doi.org/10.1016/j.neubiorev.2018.06.009

Belland, B.R., Kim, C., Dinç, E. et al. (2024*). Transfer of responsibility from scaffolding to preservice early childhood teachers learning to debug*. Education Tech Research Dev 72, 1439–1464 . https://doi.org/10.1007/s11423-024-10347-z

Bernal Quispe, Rosa Angela, Camacho Chávez, Percy, Asto Huamaní, Alina Yvón, & Salvatierra Melgar, Ángel. (2025). *Adaptación y validación de un instrumento que mide el pensamiento crítico en estudiantes de secundaria*. Revista Tribunal, 5(11), 300-310. Epub 01 de abril de 2025. https://doi.org/http://doi.org/10.59659/revistatribunal.v5i11.158

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. David McKay Company.

Brown, J. S., Collins, A., & Duguid, P. (1989). *Situated cognition and the culture of learning.* Educational Researcher, 18(1), 32–42.

Bunge, M. (1969). *La investigación científica. Su estrategia y su filosofía*. (M. Sacristán, Trad.). Barcelona: Ariel.

Carrasco Díaz, S. (2019). *Metodología de la investigación* (19a ed.). Editorial San Marcos EIR Ltda.

Castillo Cuadra, R. A. (2020). *El Pensamiento Crítico como competencia básica. Una propuesta de nuevos estándares pedagógicos*. IXTLI: Revista Latinoamericana de Filosofía de la Educación, 7(14), Article 14. https://dialnet.unirioja.es/servlet/articulo?codigo=7568883

Castillo Herrera, M. E. (2023). *Impacto de la inteligencia artificial en el proceso de enseñanza y aprendizaje en la educación secundaria*. LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades, 4(6), 515 – 530. https://doi.org/10.56712/latam.v4i6.1459

Castillo, M. E. (2023). Impacto de la inteligencia artificial en el proceso de enseñanza y aprendizaje en la educación secundaria. LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades, 4(6), 515 – 530. https://doi.org/10.56712/latam.v4i6.1459

Castro, L. A., Terrones, M. A., Duran, K. L., & Oscar, G. M. (2023). *Estrategia aprendizaje basado en proyectos para desarrollar el pensamiento crítico en estudiantes de secundaria.* Revista Arbitrada Interdisciplinaria Koinonía, 8(Supl. 2), 149-162. https://doi.org/10.35381/r.k.v8i2.2868

Chacón, J., Duran, K., Chacón, G., & Bustamante, D. (2023). *Aprendizaje basado en problemas para desarrollar el pensamiento crítico en estudiantes de secundaria*. Revista Arbitrada Interdisciplinaria Koinonía, 8(Supl. 2), 350-370. https://doi.org/10.35381/r.k.v8i2.2880

Comisión Económica para América Latina y el Caribe. (2023). *Mejorar la educación es crucial para un desarrollo social y económico inclusivo y sostenible*. CEPAL. https://repositorio.cepal.org/server/api/core/bitstreams/ea5665bd-468c-4864-ba3c-0c890651f617/content

Consejo Nacional de Educación (2022). *Problemas críticos de la educación secundaria en el Perú*. CNE. https://cdn.www.gob.pe/uploads/document/file/6172492/5445311-documento-de-trabajo-problemas-criticos-de-la-educacion-secundaria-en-el-peru.pdf?v=1712770560

Dávila Newman, G. (2006). *El razonamiento inductivo y deductivo dentro del proceso investigativo en ciencias experimentales y sociales.* Laurus, 12(Ext), 180-205. https://www.redalyc.org/articulo.oa?id=76109911

De la Cruz, G., Villalobos, J., & Morales, S. (2023). Comprensión lectora y pensamiento crítico: Procesos de análisis e interpretación en la educación secundaria. Revista Innova Educación, 5(1), 125-140. https://doi.org/10.35622/j.rie.2023.05.008

Dewey, J. (1933). *How we think: a restatement of the relation of reflective thinking to the educative process*. Boston: Heath & Co. Publishers.

Downes, S. (2008). *Connectivism and its practical application for technology education.* In M. A. Littlejohn (Ed.), Technology and higher education: An examination of new learning environments (pp. 57-73). Nova Science Publishers.

Durning, S. J., Schuwirth, L., & Artino, A. R. (2020). *Situated cognition: A theoretical framework for the health professions*. En Theories in health professions education (2nd ed., pp. 159-166). Routledge.

Ennis, R. H. (1987). *A taxonomy of critical thinking dispositions and abilities*. In J. B. Baron & R. J. Sternberg (Eds.), Teaching thinking skills: Theory and practice (pp. 9–26). W H Freeman/Times Books/ Henry Holt & Co.

Ennis, R. H. (2011). The nature of critical thinking: An outline of critical thinking dispositions and abilities. University of Illinois. https://education.illinois.edu/docs/default-source/faculty-documents/robert-ennis/thenatureofcriticalthinking_51711_000.pdf

Espinal, C. A., Tapia, A., Guerra, D. L., & Martel, L. V. (2022). *Aprendizaje colaborativo para la mejora del pensamiento crítico en estudiantes de secundaria*. *Horizontes. Revista de Investigación en Ciencias de la Educación, 6*(26), 1951–1960. https://doi.org/10.33996/revistahorizontes.v6i26.464

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report).* American Philosophical Association.

Facione, P. A. (2020). *Critical thinking: What it is and why it counts*. Insight Assessment.

Fondo de las Naciones Unidas para la Infancia (2021). *Importancia del desarrollo de habilidades transferibles en América Latina y el Caribe*. UNICEF. https://www.unicef.org/lac/media/21536/file/Importancia_Desarrollo_Habilidades_Transferibles_ALC_v.actualizada_marzo2021.pdf

Halpern, D. F. (1998). *Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring*. American Psychologist, 53(4), 449–455. https://doi.org/10.1037/0003-066X.53.4.449

Hasanah, H., & Malik, M. N. (2020). *Blended learning in improving students' critical thinking and communication skills at University*. Cypriot Journal of Educational Sciences, 15(5), 1295–1306. https://doi.org/10.18844/cjes.v15i5.5168

Hernández-Sampieri, R. & Mendoza, C. (2023). Metodología de la investigación. *Las rutas cuantitativa, cualitativa y mixta* (2ª ed.). McGraw-Hill Education.

Hobert, S. (2023). *Fostering skills with chatbot-based digital tutors - training programming skills in a field study. i-com*, 22(2), 143-159. https://doi.org/10.1515/icom-2022-0044

Humpiri, J., Humpiri, FdeM., & Mamani, E. E. (2021). Teorías científicas. Las propuestas de Popper y Kuhn sobre investigaciones científicas. Horizontes Revista de Investigación en Ciencias de la Educación, 5(17), 277-296. https://doi.org/10.33996/revistahorizontes.v5i17.171

Jitendra, A. K., & Lein, A. E. (2020). *Mathematical problem solving for students with learning disabilities and those at risk: The role of schema-based instruction*. International Journal of Educational Research, 102, Artículo 101587. https://doi.org/10.1016/j.ijer.2020.101587

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Harvard University Press.

Köksal, N., & Çöğmen, S. (2018). *Pensamiento crítico y habilidades comunicativas en estudiantes de secundaria. Revista de Educación de la Universidad de Pamukkale, 44*(44), 278-296. https://doi.org/10.9779/pauefd.422244

Kurmanbayeva, N., & Besekeeva, B. (2025). The contribution of artificial intelligence (AI) in the development of students' critical thinking. Eurasian Science Review an International Peer-Reviewed Multidisciplinary Journal, 2(Special Issue), 1959–1970. https://doi.org/10.63034/esr-320

Laun, M., & Wolff, F. (2025). Chatbots in education: hype or help? A meta-analysis. Learning and Individual Differences, 119, 102646. https://doi.org/10.1016/j.lindif.2025.102646.

Ledesma, C. L., & Sevairos, J. K. (2023). *Uso de herramientas digitales para el desarrollo del pensamiento crítico en estudiantes. Horizontes. Revista de Investigación en Ciencias de la Educación, 7*(27), 509–517. https://doi.org/10.33996/revistahorizontes.v7i27.533

Leighton, J. P., Cui, Y., & Cutumisu, M. (2021). *Key Information Processes for Thinking Critically in Data-Rich Environments. Frontiers in Education, 6*, 30. https://doi.org/10.3389/FEDUC.2021.561847/BIBTEX

Macedo Dávila, A. (2018). *Pensamiento crítico y rendimiento académico en los ingresantes del curso de Estadística I en la Facultad de Ingeniería Económica, Estadística y Ciencias Sociales.* Universidad Nacional de Ingeniería-2017.

Medina Salazar, F. (2017). *Programa critical para desarrollar pensamiento crítico en estudiantes de cuarto de secundaria de una institución educativa del Callao [Tesis de maestría, Universidad San Ignacio de Loyola].* http://repositorio.usil.edu.pe/bitstream/USIL/2997/1/2017_Medina_Programacritical-para-pensamiento-critico.pdf.

Mendoza, I.; Gómez, M. y García, I. (2016). *El desarrollo del pensamiento crítico mediante el debate asincrónico en foros virtuales en educación secundaria*. *Revista Aletheia, 8*(1), 100–115. http://dx.doi.org/10.11600/21450366.8.1aletheia.80.99

Ministerio de Educación del Perú. (2016). *Evaluación censal de estudiantes: Resultados y análisis*. MINEDU. https://www.minedu.gob.pe

Ministerio de Educación del Perú. (2017). *Rúbrica de Desempeño Docente*. MINEDU, Lima, Perú

Ministerio de Educación del Perú. (2022). *Reporte de Resultados del Monitoreo de Prácticas Escolares 2022*. MINEDU. https://repositorio.minedu.gob.pe/handle/20.500.12799/8988

Ministerio de Educación del Perú. (2023). *Resultados de la Evaluación Nacional de Estudiantes (ENE)*. MINEDU.

Moreira-Choez JS, Núñez-Naranjo AF, Carrasco-Valenzuela AC et al. (2025) *Algoritmos de aprendizaje*

automático para predecir las competencias digitales en el profesorado universitario [versión 1; revisión por pares: pendiente de revisión por pares] . F1000Research, 14 :573. https://doi.org/10.12688/f1000research.165342.1

Objetivos de Desarrollo Sostenible 4. *Las conclusiones del monitoreo del ODS 4 2015-2021: Diez desafíos para enfrentar*

Ocampo Flórez, E. O. (2017). *Modelos mentales sobre aprendizaje en estudiantes de pedagogía infantil.* https://biblioteca.clacso.edu.ar/Colombia/alianza-cinde-umz/20170925112658/EstebanOcampoFlorez.pdf

Organización de Estados Iberoamericanos. (2020). *La educación por competencias se impone en los planes de estudio de las aulas iberoamericanas*. OEI. https://oei.int/oficinas/secretaria-general/noticias/la-educacion-por-competencias-se-impone-en-los-planes-de-estudio-de-las-aulas-iberoamericanas-3/

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (2023). *El pensamiento crítico y la interdisciplinariedad son esenciales para preparar a los alumnos para el cambio climático*. UNESCO. https://www.unesco.org/es/articles/el-pensamiento-critico-y-la-interdisciplinariedad-son-esenciales-para-preparar-los-alumnos-para-el.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (2024). *Hace un llamamiento urgente para un uso adecuado de la tecnología en la educación.* UNESCO. https://www.unesco.org/es/articles/la-unesco-hace-un-llamamiento-urgente-para-un-uso-adecuado-de-la-tecnologia-en-la-educacion

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2024). Marco de competencias para docentes en materia de IA. https://unesdoc.unesco.org/ark:/48223/pf0000393813

Organización para la Cooperación y el Desarrollo Económicos (2023). *PISA 2022 Results: Creative Thinking*. OECD Publishing.

Orhan, A. & Çeviker-Ay, Ş. (2022). *Developing the critical thinking skill test for high school students: a validity and reliability study*. International Journal of Psychology and Educational Studies, 9(1), 132-44. https://dx.doi.org/10.52380/ijpes.2022.9.1.561

Osmanović, J., & Maksimović, J. (2022). *Quasi-Experimental Research as An Epistemological-Methodological Approach in Education Research. International Journal of Cognitive Research in Science, Engineering and Education (IJCRSEE), 10*(3), 177–183. https://doi.org/10.23947/2334-8496-2022-10-3-177-183

Paas, F., Renkl, A., & Kirschner, P. A. (2004). *Effects of the worked-example principle on efficiency and effectiveness of learning revisited*. Educational Technology Research and Development, 52(4), 89–101

Padilla, D. (2023). *GPT-3 y enseñanza de la educación superior: ¿tecnofobias y tecnofilias? Kinesis.* Revista Veracruzana de Investigación Docente, 8(8), 84-99,

Paul, R., & Elder, L. (2003). *La mini-guía para el pensamiento crítico, conceptos y herramientas*. Dillon Beach, California: Fundación para el pensamiento. crítico. https://didacticadelaeducacionsuperioruacj.wordpress.com/wp-content/uploads/2017/05/guc3aca-para-el-pensamiento-crc3actico.pdf

Pavlou, V. (2020). *Art Technology Integration: Digital Storytellying as a Transformative Pedagogy in Primary Education*. International Journal of Art & Design Education, *39*(1), 195–210. https://doi.org/10.1111/JADE.12254

Paz, J. S., Molina, E. C., & Sánchez, L. P. (2010). *Pensamiento crítico y capacidad intelectual.* Faisca: revista de altas capacidades, 15(17), 92-110. https://dialnet.unirioja.es/servlet/articulo?codigo=3548104

Pérez Pérez, J. A., Holguín Holguín, M. S., Posligua Anchundia, M. L., & Párraga Flores, L. L. (2025). *El pensamiento crítico según el test de Watson-Glaser en estudiantes del cuarto nivel de la carrera de Educación Básica: Critical thinking according to the Watson-Glaser test in students of the fourth level of the Basic Education career.* Revista Cognosis. ISSN 2588-0578, 10(1), 42–52. https://doi.org/10.33936/cognosis.v10i1.6614

Programa de Promoción de la Reforma Educativa en América Latina y el Caribe (2022). *Informe sobre la calidad docente en América Latina. Diálogo Interamericano. Puig, B., Blanco, P., & Pérez, J. J. (2021). Fake news" or real science? Critical thinking to assess information on COVID-19. Frontiers in Education, 6*, 646909. https://doi.org/10.3389/feduc.2021.646909

Quintana, R., & Correnti, R. (2019). *The right to argue: Teaching and assessing everyday argumentation skills.* Journal of Further and Higher Education, 43(8), 1133-1151. https://doi.org/10.1080/0309877X.2018.1450967

Ramos, C. (2021) *Editorial: Diseños de investigación experimental*. Enero – Junio 2021 http://dx.doi.org/10.33210/ca.v10i1.356

Reyes, E. (2022). *Metodología de la Investigación científica*. Page Publishing Inc.

Romaní, B. S., Martell, M. A., Medina, R. J., & Criado, Y. V. (2021). *Percepción del pensamiento crítico en estudiantes de una institución educativa de Lima, 2021. IGOBERNANZA*, *4*(16), 62–87. https://doi.org/10.47865/igob.vol4.2021.151

Rumelhart, D. E. (1980). *Schemata: The building blocks of cognition. En R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education* (pp. 33–58). Lawrence Erlbaum Associates.

Saavedra, L. (2024). *Pensamiento crítico en estudiantes de educación básica regular. Horizontes.* Revista de Investigación en Ciencias de la Educación, *8*(33), 809–819. https://doi.org/10.33996/revistahorizontes.v8i33.764

Sánchez, R. A. (2015). Prueba de Wilcoxon-Mann-Whitney: mitos y realidades. Revista Mexicana de Endocrinología, Metabolismo y Nutrición, 2, 18-21. https://biblat.unam.mx/hevila/Revistamexicanadeendocrinologiametabolismo%26nutricion/2015/vol2/no1/3.pdf

Siemens, G. (2005). *Connectivism: A learning theory for the digital age.* International Journal of Instructional Technology and Distance Learning, 2(1), 3–10.

Sperling, K., Stenberg, C. J., McGrath, C., Åkerfeldt, A., Heintz, F., & Stenliden, L. (2024). In search of artificial intelligence (AI) literacy in teacher education: A scoping review. *Computers and Education Open, 6*, 100169. https://doi.org/10.1016/j.caeo.2024.100169

Sweller, J. (1980). Transfer effects in a problem solving context. The Quarterly Journal of Experimental Psychology, 32(2), 233-239.

Tantajulca Chávez, E. I. (2025). El pensamiento crítico y su importancia en la educación básica latinoamericana. *Revista de Ciencias Humanísticas y Sociales (ReHuSo), 10*(1), 10-21. https://doi.org/10.33936/rehuso.v10i1.6494

Tobón, S. (2017). *Evaluación socioformativa: Estrategias e instrumentos para la formación integral*. Ecoe Ediciones.

Twabu, K. (2025). Enhancing the cognitive load theory and multimedia learning framework with AI insight. Discover Education, 4(1), 160. https://link.springer.com/article/10.1007/s44217-025-00592-6

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

Vivanco Enríquez, Jesús (2026). Creatividad y Pensamiento Crítico en la Educación Peruana: una Lectura Inicial desde el Currículo y las Evaluaciones Internacionales PISA 2022. Vol. 32 Núm. 1 (2026): Educación: Revista de la Facultad de Ciencias de la Educación. DOI: https://doi.org/10.33539/educacion.2026.v32n1.3637

Watson, G., y Glaser, M. (1994). *Watson-Glaser critical thinking appraisal form S manual.* The Psychological Corporation.

Watson, G., & Glaser, E. (2012). *Watson-Glaser Valoración del pensamiento crítico*. Editorial Reino Unido CW.

Watson, G., & Glaser, E. M. (2009), Watson-Glaser Critical Thinking Appraisal Short Form, USA: Pearson Education, Inc.

Wood, D., Bruner, J. S., & Ross, G. (1976). *The role of tutoring in problem solving*. Journal of Child Psychology and Psychiatry, 17(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Yim, IHY, Su, J. (2025) *Herramientas de aprendizaje de inteligencia artificial (IA) en la educación primaria y secundaria*: *Una revisión exploratoria*. J. Comput. Educ. 12 , 93–131. https://doi.org/10.1007/s40692-023-00304-9