Journal of Education and Learning Reviews, 3 (6): November-December 2026. ISSN: 3057-0387
Website: https://so19.tci-thaijo.org/index.php/JELS/about
Article ID: e3979

Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development

Aniza C. Cargason
Saint Columban College, Pagadian City, Philippines
E-mail: aniza.cargason@sccpag.edu.ph ORCID ID: https://orcid.org/0009-0000-0949-7226

Shaniz F. Dumam-ag
Saint Columban College, Pagadian City, Philippines
E-mail: shaniz.dumam-ag@sccpag.edu.ph *ORCID ID: https://orcid.org/0009-0007-0203-0990*

Jehan U. Amol
Saint Columban College, Pagadian City, Philippines
E-mail: jehan.amol@sccpag.edu.ph ORCID ID: https://orcid.org/0009-0009-2623-8598

Samcedar M. Dirampatun
Saint Columban College, Pagadian City, Philippines
E-mail: samcedar.dirampatun@sccpag.edu.ph ORCID ID: https://orcid.org/0009-0002-2667-0515

Matthew Sabasales
Saint Columban College, Pagadian City, Philippines
E-mail: matthew.sabasales@deped.gov.ph, ORCID ID: https://orcid.org/0009-0003-0430-3121

Received 14/09/2026	Revised 04/10/2026	Accepted 30/09/2026

Abstract

Background and Aim: Evidence on generative artificial intelligence (AI) in secondary science education remains limited. This systematic review examined students' conceptual understanding and engagement, teachers' pedagogical strategies and challenges, and students' critical thinking and scientific inquiry.

Materials and Methods: Following PRISMA 2020, ERIC, Scopus, and Lens.org were searched for peer-reviewed empirical studies published between 2022 and 2026. Two authors independently screened 187 unique records (Cohen's $\kappa = .78$), identifying 20 eligible studies. Methodological quality was assessed using the Mixed Methods Appraisal Tool. Framework synthesis integrated findings, informed by Technological Pedagogical Content Knowledge (TPACK).

Results: Seven studies assessed AI capabilities rather than classroom implementation. Only two compared learning outcomes across conditions; none employed a randomized controlled design. Student perceptions and engagement were generally positive, whereas learning outcomes were mixed. A chatbot supporting self-regulated learning outperformed a rule-based chatbot, but ChatGPT-4o yielded smaller gains than peer collaboration. Teachers reported concerns regarding content accuracy, assessment reliability, and instructional readiness. Generative AI matched or exceeded student performance on structured examinations but demonstrated lower grading reliability than human assessors. Evidence concerning critical thinking was particularly limited. Richer reasoning was reported with dialogue-oriented tools or teacher guidance, whereas answer-seeking use of general-purpose tools was associated with weaker reasoning.

Conclusion: Pedagogical structuring was recurrently associated with productive AI use, although causal inference remains limited. A provisional model warrants empirical testing. Priorities include structured tasks, science-specific teacher development, human assessment oversight, explicit AI literacy instruction, school governance, and controlled comparisons of structured and unstructured use across secondary science classrooms.

Keywords: Generative Artificial Intelligence; High School Science Education; Systematic Review; Technological Pedagogical Content Knowledge; Critical Thinking

Introduction

Generative artificial intelligence (generative AI) refers to systems that produce new text, images, code, or other content in response to natural-language prompts; in education, the most widely used are chatbots built on large language models, such as ChatGPT, Gemini, and Claude. Since the public release of ChatGPT in November 2022, these tools have entered schools faster than research has documented their educational use. Reviews of the broader literature show that the available evidence is

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

concentrated in higher education. Lee et al. (2025), for example, found that 69.7% of 92 studies on ChatGPT in education published between November 2022 and March 2024 were conducted in higher education settings and only 3.4% in primary education, and Yusuf et al. (2024) likewise reported comparatively few studies situated in K-12 settings. Reviews that do focus on secondary schooling describe implementations that are mostly teacher-initiated and confined to isolated or experimental settings (Pérez-Montesdeoca et al., 2026). As a result, much of what is currently known about learning with generative AI derives from adult learners in university contexts.

High school science warrants separate examination for three reasons. First, high school students are adolescents whose capacities for self-regulation and for monitoring their own reasoning are still developing (Steinberg, 2005), so the same tool may be used differently, and with different consequences, by a secondary student than by a university student. Second, in many education systems high school is the stage at which science instruction becomes differentiated into disciplines such as physics, chemistry, and biology, relies increasingly on abstract models and quantitative problem solving, and is assessed through high-stakes examinations, which distinguishes it from the more general science of earlier grades. Third, science education aims not only at the acquisition of content knowledge but also at the development of scientific practices such as explanation, argumentation, modeling, and inquiry (Osborne, 2014). Generative AI can support these practices by offering alternative explanations, generating practice problems, and providing immediate feedback, but it can also perform them on students' behalf, reducing opportunities for independent reasoning. Cheung et al. (2025) accordingly identified potential conflicts between the capabilities of generative AI and the goals of science education, including science understanding and epistemic literacy. These features make findings from higher education, from middle school, or from K-12 education as a whole difficult to transfer directly to the high school science classroom.

How these possibilities play out depends heavily on decisions made beyond the individual learner. Teachers decide whether, when, and for what purposes generative AI enters a lesson, while schools and education systems set the conditions under which it does so, including professional development, assessment and academic-integrity policies, data protection for students who are minors, equitable access to tools, and supporting infrastructure (Miao & Holmes, 2023). The Technological Pedagogical Content Knowledge (TPACK) framework (Mishra & Koehler, 2006) offers a way of understanding this dependence, holding that effective technology integration arises from the interplay of teachers' technological, pedagogical, and content knowledge rather than from access to the technology alone. Applied to generative AI in high school science, TPACK implies that student outcomes cannot be understood separately from the instructional strategies teachers adopt and the challenges they encounter, which is why this review treats student learning, teacher practice, and higher-order reasoning as connected dimensions.

Several reviews have begun to consolidate this literature, but each leaves part of the picture unaddressed. Fock and Siller's (2025) scoping review of 183 publications on generative AI in secondary STEM education found that research has centered on the systems themselves: roughly 35% of publications analyzed AI performance, 22% developed new AI tools, and about half of the 139 empirical studies examined AI-generated content rather than use by students or teachers, while outcomes such as critical thinking received little attention. Cheung et al. (2025) reviewed 22 empirical studies of generative AI in science education, only six of them in secondary education, and found that content knowledge acquisition was prioritized over disciplinary practices and that epistemic and social learning goals remained under-investigated. Gunsaldi et al. (2025) synthesized 12 studies but restricted their scope to middle school science, and Azizi et al. (2026) reviewed 40 studies of artificial intelligence in physics education, most involving undergraduate participants, noting that teacher-focused and long-term studies remained limited. Pérez-Montesdeoca et al. (2026) examined secondary education but across all school subjects. Taken together, these reviews establish that school-level evidence is scarce, that much of it evaluates AI capability rather than classroom use, and that higher-order reasoning remains under-examined. None of them, however, combines a focus on high school science with an appraisal of the methodological quality of the included studies, a joint synthesis of student outcomes,

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

teacher practices, and critical thinking, and a separation of evidence on classroom use from evidence on what AI systems can do on their own.

This review addresses that gap. It systematically synthesizes empirical studies published between 2022 and 2026 on the use of generative AI in high school science education, defined as grades 9 to 12 or their national equivalents, across three dimensions: student conceptual understanding and engagement, the pedagogical strategies and challenges reported by teachers, and the development of critical thinking and scientific inquiry. Because the available studies differ widely in design, and many are small-scale, descriptive, or qualitative, the review aims to identify the effects, associations, and recurring conditions reported across studies rather than to estimate the causal effect of generative AI. TPACK serves as the interpretive lens for explaining why reported outcomes vary across classrooms. The review contributes a quality-appraised account of the high school science evidence, a distinction between classroom-use and AI-capability evidence, and a provisional conceptual model of the conditions associated with more and less productive uses of generative AI, intended to inform science teachers, school leaders, and future research.

Objectives

This review aimed to systematically identify, appraise, and synthesize empirical evidence published from 2022 to 2026 on the educational use of generative AI in high school science education, with particular attention to student learning outcomes, pedagogical integration and implementation challenges, and higher-order scientific reasoning.

Specifically, this study sought to answer the following questions:

1. What effects on student conceptual understanding and engagement, or associations with these outcomes, have been reported for generative AI use in high school science education?
2. What pedagogical strategies do high school science teachers adopt when integrating generative AI, and what challenges and conditions, including the reliability of generative AI for instruction and assessment, affect its implementation?
3. What evidence has been reported on the relationship between generative AI use and the development of high school students' critical thinking and scientific inquiry skills?

Conceptual Framework

This review is anchored in the Technological Pedagogical Content Knowledge (TPACK) framework of Mishra and Koehler (2006), which builds on Shulman's (1986) concept of pedagogical content knowledge. TPACK holds that effective use of technology in teaching comes not from technical skill alone but from the integration of three kinds of teacher knowledge: knowledge of the technology, of pedagogy, and of the subject matter. In a high school science classroom, for example, a teacher using generative AI must be able to judge whether the tool's explanation of a scientific concept is accurate, decide how students should interact with the tool, and connect both to how the concept is best taught. This explains why the same generative AI tool may produce different outcomes in different classrooms.

Because TPACK describes teacher knowledge rather than student learning, two complementary perspectives are used to interpret student outcomes. Self-regulated learning theory (Zimmerman, 2002) describes how students plan, monitor, and evaluate their own learning, processes that generative AI may support when its use is structured. Cognitive offloading (Risko & Gilbert, 2016) describes the use of external tools to reduce mental effort, which can free attention for higher-order thinking but can also replace the very reasoning a task is meant to develop.

These perspectives are organized into an input–process–output model (Figure 1), which serves as the analytical framework of this review rather than as a theory in its own right. The input component refers to the conditions under which generative AI enters the classroom: the tool used; the subject, grade level, and country; and the teacher's knowledge and the school's conditions. The process component refers to how generative AI is integrated into instruction: the role given to the tool, the degree of pedagogical scaffolding, and the strategies teachers adopt and the challenges they encounter. This component corresponds to the second research question. The output component refers to student

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Journal of Education and Learning Reviews, 3 (6): November-December 2026. ISSN: 3057-0387
Website: https://so19.tci-thaijo.org/index.php/JELS/about
Article ID: e3979

outcomes, namely conceptual understanding and engagement (RQ1) and critical thinking and scientific inquiry (RQ3).

The framework guided both data extraction and synthesis. Each included study was coded according to its input, process, and output characteristics and the TPACK domains its findings informed (see Methodology), and the resulting patterns were interpreted through TPACK and the two student-level perspectives. Student outcomes are therefore treated not as automatic results of using generative AI but as dependent on how the tool is integrated into instruction.

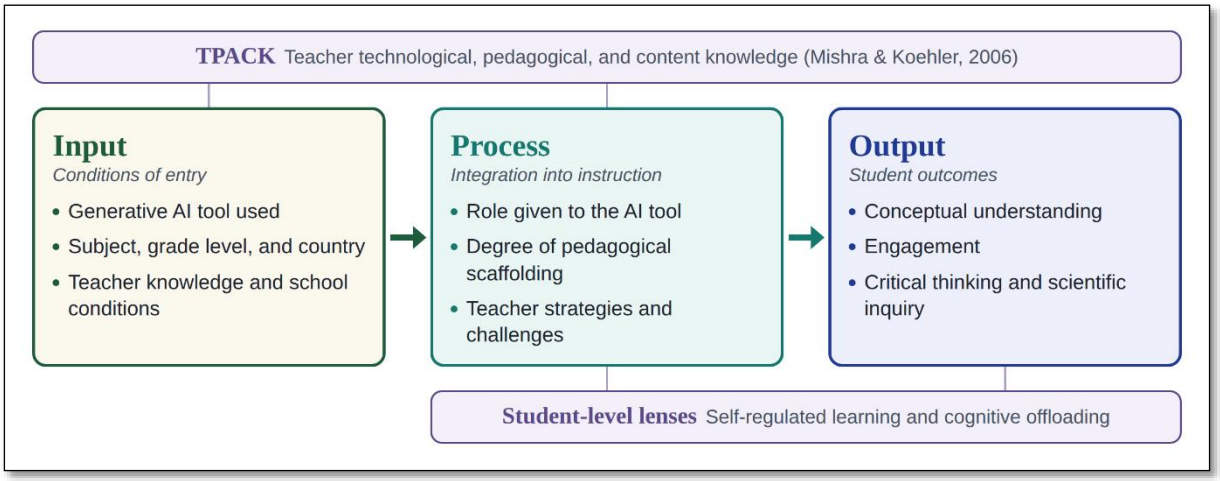


Figure 1. Analytical Framework of the Review

Literature Review

This section reviews background literature that lies outside the final review sample. None of the studies discussed here met the review's eligibility criteria: they involved other grade levels, subjects other than a specific science discipline, or mixed secondary and university populations, or they were reviews rather than primary studies. The purpose of this section is therefore not to anticipate the review's findings, which are reported in the Results, but to identify the mechanisms that earlier work suggests may explain why generative AI produces mixed outcomes in science classrooms. Four such mechanisms are examined in turn: the gap between adoption and structured guidance, the mediating role of teacher knowledge, ethical and equity conditions, and the difference between structured and unrestricted student access.

Adoption Ahead of Structured Guidance

In many schools, student and teacher use of generative AI has preceded structured guidance on how to use it. In the Philippines, Garcia et al. (2025) surveyed 899 junior high school students and 26 science teachers and found that both groups regarded ChatGPT favorably for classroom engagement and the comprehension of scientific concepts. Among 525 Filipino senior high school students, Co et al. (2026) found that 85.7% used ChatGPT but had learned to do so mainly informally and from peers rather than through teacher guidance; AI literacy was a modest but significant predictor of academic performance, explaining roughly 7% of the variance in grades, with its awareness and ethics dimensions contributing more than technical usage. On the teacher side, Ajose et al. (2026) found that most of the 57 K-12 STEM educators they examined were aware of ChatGPT, but few used it extensively and most rated its impact on STEM teaching as minimal, regardless of years of teaching experience. Marzano's (2026) review of 197 studies of generative AI at the K-12 level likewise identified a notable lack of concrete studies and classroom experiments relative to theoretical and policy-oriented work. These studies are cross-sectional and perception-based, and they span grade levels and subjects beyond high school science, so they cannot show how generative AI affects learning. They do, however, suggest a first mechanism behind mixed outcomes: when students use generative AI before they have been taught

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

to evaluate its output, the quality of that use depends on informal habits rather than on instruction. If structured guidance is to come from anywhere, it is most likely to come from teachers, which places teacher knowledge at the center of the next body of work.

Teacher Knowledge as a Mediating Condition

Studies of teachers' perspectives suggest that what teachers know shapes how productively they can use generative AI. Shankar et al. (2025) analyzed open-ended responses from 67 Indian high school educators and found that they most readily identified strengths such as personalized learning, content enrichment, and interactive learning experiences. Jang and Choi (2025), interviewing ten Korean physics educators from secondary schools and universities, found that participants saw potential for personalized, student-directed inquiry but stressed that realizing it would require curricular and infrastructural change and stronger teacher digital competencies. A larger survey of 320 Thai physics instructors, most of them secondary school teachers, identified incorrect or incomplete AI-generated information as the most frequently reported challenge (68.4%), followed by concern for students' analytical thinking (58.7%); insufficient technical knowledge was the main barrier among non-users, and instructors who paid for premium tools were far more likely than free-tool users to request guidance on prompt writing (Wattanakasiwich et al., 2025). Read through TPACK, these findings point to two specific demands: detecting errors in AI-generated science content requires technological content knowledge, and designing tasks in which students reason with the tool rather than simply retrieve answers from it requires technological pedagogical knowledge. The evidence nevertheless has clear limits. It relies mainly on self-report, sometimes combines secondary and university educators, and comes largely from physics, leaving chemistry and biology teachers comparatively under-represented.

Ethical and Equity Conditions

A third set of studies shows that the conditions under which generative AI is adopted are not equal across students and teachers. Both the students and the science teachers surveyed by Garcia et al. (2025) raised concerns about bias, privacy, and the ethical implications of generative AI in schools. In a grounded theory study of seven primary and secondary STEM educators in Australia, Tipple et al. (2026) found that teachers worried that biased AI outputs could reinforce gender disparities in STEM, even as they saw the technology's capacity to scaffold confidence and tailor learning as a way of engaging girls; strategies for mitigating bias were recognized mainly by teachers with more generative AI experience. Access is also uneven: most Thai physics instructors relied exclusively on free versions of generative AI tools, and half of users reported limitations in Thai-language processing (Wattanakasiwich et al., 2025). These conditions suggest a third mechanism: the same tool may offer different benefits depending on students' language, gender, and access to paid versions, and on whether their teachers can recognize and correct biased output. As with the teacher-knowledge studies, this evidence is largely survey-based or qualitative and drawn from small or relatively advantaged samples.

Structured Versus Unrestricted Access

The most rigorous evidence on learning effects comes from experimental studies in adjacent populations. In a field experiment with nearly 1,000 high school mathematics students, Bastani et al. (2025) found that access to GPT-4 while solving practice problems improved performance substantially, but once the tool was removed, students who had used an unrestricted chatbot performed 17% worse than students who never had access; a version designed with learning safeguards largely mitigated this harm. Fütterer et al. (2026) conducted a randomized controlled trial with 371 German students in grades 7 to 9 across physics and English lessons, comparing generative AI support designed to foster self-regulated learning with a standard ChatGPT condition. A utility-value version produced more favorable development of perceived utility value than a cognitive-strategy version, but neither showed a statistically significant advantage over the standard condition on effort, domain-specific knowledge, or elaboration-based strategy use. Neither study falls within the scope of the present review, the first because it concerns mathematics and the second because its sample was drawn from grades 7 to 9, but together they illustrate a fourth mechanism, consistent with the concept of cognitive offloading: unrestricted access can improve immediate performance while reducing independent learning, and adding structure can limit this harm without guaranteeing gains in knowledge. They also caution that rigorous designs may find smaller effects than those suggested by small, uncontrolled studies.

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Synthesis and Implications for the Review

Summing everything up, this background literature suggests four mechanisms that may explain why generative AI yields mixed outcomes in science classrooms: adoption that runs ahead of structured guidance and student AI literacy; the extent to which teachers can detect errors and design productive interactions; ethical and equity conditions that shape who benefits; and the difference between structured and unrestricted student access. These mechanisms correspond closely to the input, process, and output components of the review's analytical framework. The same literature also reveals limitations that shaped the design of this review. Because much of it mixes grade levels, subjects, and educator populations, the review restricted its scope to high school science and recorded the grade level and discipline of each included study, making any imbalance across physics, chemistry, and biology visible. Because the designs are predominantly cross-sectional, qualitative, or small-scale, the review appraised methodological quality using the MMAT rather than treating all findings as equivalent. Because some studies assess what generative AI can do rather than what students learn, the review classified each study by type of evidence. Finally, because the contrast between structured and unrestricted use emerges from adjacent populations, the review coded the role assigned to generative AI and the degree of pedagogical structuring in each study to examine whether the same pattern holds in high school science.

Methodology

Review Design

This study is a systematic review reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement (Page et al., 2021). A systematic rather than scoping design was chosen because the aim was to appraise and synthesize the empirical evidence relevant to three defined questions, not only to map the breadth of the literature. The review protocol was not registered in advance. The full search strings, the reports excluded at full text with reasons, and the reports that could not be retrieved are provided in Supplementary Material S1.

Information Sources and Search Strategy

Three databases were searched: ERIC, for its coverage of education research; Scopus, as a large multidisciplinary index of peer-reviewed journals; and Lens.org, an open scholarly aggregator that draws on sources such as Crossref, PubMed, and OpenAlex and therefore captures journals not indexed in the other two. All three databases were searched on 12 September 2026. The search combined three concept blocks with the Boolean operator AND: generative AI terms (such as generative AI, ChatGPT, large language models, and chatbots), secondary education terms (such as high school, secondary school, and K-12), and science subject terms (such as science, physics, chemistry, biology, and STEM). Terms within each block were combined with OR, adapted to the syntax of each database, and applied to titles, abstracts, and keywords. Results were limited to peer-reviewed journal articles in English published from 2022 to 2026, the start year reflecting the public release of ChatGPT in November 2022. In Scopus, results were additionally limited to open-access publications; the implications of this restriction are considered in the Limitations. The searches returned 84 records from ERIC, 21 from Scopus, and 83 from Lens.org, a total of 188. Records were imported into Rayyan (Ouzzani et al., 2016) for de-duplication and record management, and one duplicate was removed. No supplementary searching, such as citation or reference-list searching, was conducted.

Eligibility Criteria

Studies were eligible if they met all of the inclusion criteria in Table 1 and none of the exclusion criteria.

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Table 1 *Eligibility Criteria*

Criterion	Inclusion	Exclusion
Population	Students in grades 9 to 12 or their national equivalents; in-service science teachers teaching these grades	Exclusively primary, middle school, or higher education populations; pre-service teachers only
Phenomenon	Use, perceived use, or evaluation of a generative AI tool, defined as a system that generates new text, images, or other content in response to prompts (e.g., ChatGPT, Gemini, Claude, or a custom chatbot built on a large language model)	Rule-based chatbots or non-generative AI, except as a comparison condition
Context	A natural science subject: physics, chemistry, biology, earth or environmental science, or general science	Non-science subjects, or broad STEM settings without separately reported science findings
Outcomes	Findings relevant to at least one research question	No findings relevant to the research questions
Design	Empirical primary studies with systematically collected data: quantitative, qualitative, or mixed methods, including studies that evaluated generative AI on high school science tasks	Reviews, conceptual papers, editorials, and opinion pieces
Publication	Peer-reviewed journal articles in English, 2022–2026	Conference abstracts, theses, preprints, and non-English publications

Three operational rules were applied. First, because grade structures differ across education systems, educational level was mapped to grades 9 to 12 by national designation and typical age (approximately 14 to 18 years); for example, grades 9 to 12 correspond to the final two years of junior high school and the two years of senior high school in the Philippines, Secondary 3 to Secondary 6 in Hong Kong, and Junior Secondary 3 and Senior Secondary 1 to 3 in Nigeria. Second, studies with mixed-level samples were included only if results for grades 9 to 12 were reported separately or if more than half of the participants were in these grades; teacher samples combining secondary and university educators were included only if secondary results were reported separately. Third, studies that evaluated the performance of generative AI on tasks drawn from high school science curricula or assessments, such as national examinations, olympiad problems for secondary students, or high school students' written work, were eligible as evidence on AI capability even when no students used the tool, and were analyzed separately from studies of classroom use. The grade level, education system, and science discipline of every included study are reported in Table 2.

Study Selection

Before screening, the collated records were organized with the help of an automated keyword-matching procedure and a generative AI assistant, which proposed provisional classifications against the eligibility criteria. These provisional classifications were not treated as eligibility decisions. All eligibility decisions were made by the first and second authors, who screened every record independently, without access to the provisional classifications or to each other's decisions.

Screening took place in two stages. At the title-and-abstract stage, each reviewer classified all unique records as include, exclude, or unsure against the criteria in Table 1, recording a single primary reason for each exclusion. Any record classified as include or unsure by either reviewer was carried forward to full-text assessment, so that records whose eligibility could not be determined from the title and abstract, such as those with mixed-level samples or subjects described only as STEM, were resolved from the full text rather than excluded. Full texts were sought through open-access sources and the authors' institutional access; reports that could not be obtained in this way were recorded as reports not retrieved, in accordance with PRISMA 2020. At the full-text stage, both reviewers independently

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Journal of Education and Learning Reviews, 3 (6): November-December 2026. ISSN: 3057-0387
Website: https://so19.tci-thaijo.org/index.php/JELS/about
Article ID: e3979

assessed each retrieved report against the same criteria and recorded one primary reason for each exclusion. At both stages, disagreements were resolved through discussion, with the third author adjudicating where agreement could not be reached. Agreement between the two reviewers was calculated separately for each stage using Cohen's kappa and interpreted following Landis and Koch (1977). The outcomes of each stage are reported in the Results and in the PRISMA flow diagram (Figure 2).

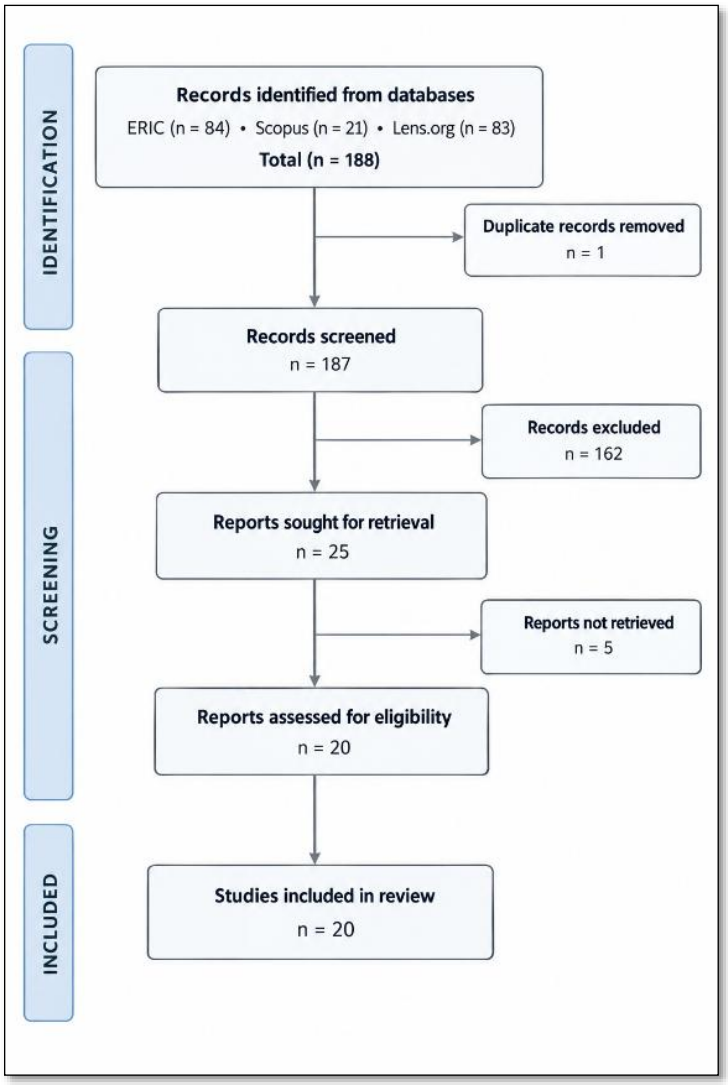


Figure 2. PRISMA 2020 flow diagram of study identification, screening, and inclusion.

Full-text copies could not be obtained for five of these 25 records due to publisher paywall restrictions and the absence of an accessible open-access or institutionally available version; these five records were consequently excluded, consistent with PRISMA's provision for reporting inaccessibility as a formal exclusion reason. The remaining 20 records were then independently assessed for full-text eligibility by both reviewers, with disagreements again resolved through discussion. Inter-rater agreement across the title/abstract and full-text screening stages, calculated using Cohen's Kappa, was .78, indicating substantial agreement (Landis & Koch, 1977). These 20 records, for which full text was successfully retrieved and eligibility confirmed by both reviewers, constitute the final sample included in this systematic review.

Data Extraction

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

The first and second authors independently extracted data from each study using a structured form and compared their entries item by item, resolving discrepancies by discussion with reference to the full text and, where needed, with the third author. The form recorded each study's bibliographic details; country and education system; grade level and age; science discipline; sample size and characteristics; generative AI tool, model, and version; design, duration, and comparison condition; data collection instruments; outcome measures; and principal findings, including effect estimates, test statistics, and reliability coefficients where these were reported, or "not reported" where they were not. In line with the analytical framework, each study was also coded for (a) its input, process, and output characteristics, including the role assigned to generative AI and the degree of pedagogical structuring of student–AI interaction (scaffolded or open-ended); (b) the TPACK domain or domains its findings informed; (c) the research question or questions to which it contributed; and (d) the type of evidence it provided. Three evidence types were distinguished: AI capability, meaning the performance of generative AI on science tasks, examinations, or grading without student use; AI-assisted student performance, meaning how students performed while using generative AI; and AI-mediated learning outcomes, meaning changes in students' understanding, engagement, or reasoning associated with AI-supported instruction.

Quality Appraisal

Methodological quality was appraised with the Mixed Methods Appraisal Tool, version 2018 (Hong et al., 2018), which accommodates the qualitative, quantitative randomized, quantitative non-randomized, quantitative descriptive, and mixed-methods designs represented in the sample. Each study was first checked against the two MMAT screening questions, then assigned to a design category and rated on the five criteria for that category as yes, no, or can't tell. For studies that evaluated generative AI rather than human participants, the quantitative descriptive criteria were applied to the tasks or responses sampled, such as examination items or graded essays. The first and second authors appraised each study independently, and disagreements were resolved through discussion, with the third author consulted where needed. Consistent with MMAT guidance, no overall score was calculated and no study was excluded on the basis of quality; criterion-level ratings for every study are reported in Table 3. Quality informed the synthesis in two ways. Findings from studies meeting fewer criteria were reported in more cautious terms and were not used on their own to support a conclusion. In addition, as a sensitivity check, the direction of findings for each research question was re-examined after setting aside studies that met fewer than three of the five criteria.

Data Synthesis

Because the included studies differed in design, outcomes, and measures, and few reported comparable effect estimates, meta-analysis was not appropriate. The review instead used a framework synthesis, reported in line with the Synthesis Without Meta-analysis (SWiM) guideline (Campbell et al., 2020), in four steps. First, studies were grouped by research question using the deductive codes of the analytical framework; a study with findings relevant to more than one question contributed to each and therefore appears in more than one results table, but is counted only once in the total sample. Second, within each research question, findings were grouped by outcome construct, as defined in the Objectives, and by evidence type, so that AI capability studies were synthesized separately from studies of student use. Third, for quantitative findings, the direction of each reported effect or association (positive, null, mixed, or negative) was recorded and tabulated; this count describes the distribution of findings and is not treated as an estimate of effect size. Qualitative findings were analyzed thematically (Thomas & Harden, 2008): the first and second authors coded the reported findings and developed descriptive themes inductively within each framework component, then compared and reconciled their themes. Fourth, contradictory findings were compared rather than averaged. For claims about effects, greater interpretive weight was given to controlled comparisons and to studies meeting more MMAT criteria; for claims about how and why generative AI was used, qualitative studies were given particular weight. TPACK and the student-level perspectives were then used to interpret the resulting patterns rather than as hypotheses to be tested.

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Journal of Education and Learning Reviews, 3 (6): November-December 2026. ISSN: 3057-0387
Website: https://so19.tci-thaijo.org/index.php/JELS/about
Article ID: e3979

Results

Characteristics of the Included Studies

The 20 included studies were published between 2023 and 2026, most of them in 2025 (n = 9) and 2024 (n = 6) (Table 2). They were conducted in 12 education systems: mainland China (n = 4); Germany, Hong Kong, Indonesia, Sweden, and the United States (n = 2 each); and Australia, Nigeria, South Africa, South Korea, Spain, and Taiwan (n = 1 each). Physics was the most frequently studied discipline (n = 9), followed by studies spanning several sciences or general science (n = 6), chemistry and environmental science (n = 2 each), and biology (n = 1). ChatGPT, in successive versions, was the most common tool; four studies used custom chatbots built on large language models, and one compared ChatGPT Plus, Gemini, and Claude 3.5 Sonnet.

The evidence differed considerably in type. Seven studies evaluated the capability of generative AI itself, by having it sit examinations, solve olympiad or concept-inventory problems, or score students' written work. Three examined teachers' perceptions and practices, four examined students' perceptions of or attitudes toward generative AI, and four examined how students performed or reasoned while using it. Only two studies measured learning outcomes associated with AI-supported instruction, both quasi-experimental comparisons of intact classes (Ng et al., 2024; Tong et al., 2025), and no randomized controlled trial was identified. Samples in studies with human participants ranged from 3 (Rappa et al., 2026) to 453 (Valeri et al., 2025), and interventions, where reported, ranged from a single laboratory session to five weeks of lessons. Seven studies contributed findings to RQ1, fourteen to RQ2, and five to RQ3; because several studies contributed to more than one question, these counts sum to more than 20.

Table 2 *Characteristics of the Included Studies*

Study	Country	Participants or materials (n)	Discipline	Design	Generative AI tool	Evidence type	RQ
Adeyele & Ramnarain (2024)	Nigeria	In-service science teachers, junior and senior secondary (15)	General science	Qualitative	ChatGPT	Teacher perception and practice	2
Bitzenbauer (2023)	Germany	Grade 12 students (53)	Physics	One-group pretest–posttest	ChatGPT	Student perception	1
Fadillah et al. (2024)	Indonesia	Upper-secondary students (334)	Physics	Cross-sectional survey	ChatGPT	Student perception	1, 3
Fekets (2025)	Taiwan	Upper-secondary students (37)	Physics	Mixed methods (usability pilot)	Custom chatbot (Botpress with OpenAI API)	Student perception	1, 2
Fernández et al. (2024)	Spain	University entrance examinations taken at the end of secondary school (8)	Chemistry	AI evaluation	ChatGPT-4	AI capability	2
Jere (2025)	South Africa	Grade 12 national examination paper (1)	Chemistry	AI evaluation	ChatGPT Plus, Gemini, Claude 3.5 Sonnet	AI capability	2
Jho & Ha (2024)	South Korea	Upper-secondary students' written arguments (180)	Argumentation (science and non-science)	AI evaluation	ChatGPT-3.5	AI capability	2

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Study	Country	Participants or materials (n)	Discipline	Design	Generative AI tool	Evidence type	RQ
Kilde-Westberg et al. (2025)	Sweden	Upper-secondary physics students (7 laboratory groups)	Physics	Qualitative case study	ChatGPT (GPT-3.5)	AI-assisted student performance	2, 3
Ng et al. (2024)	Hong Kong	Secondary 4 (grade 10) students (74)	Physics	Quasi-experimental	Custom chatbot built on ChatGPT, compared with a rule-based chatbot	AI-mediated learning outcomes	1
Nguyen et al. (2025)	United States	Design team of high school students, teachers, and other educators (14)	Environmental science	Qualitative, participatory design	Custom chatbots built on large language models	Teacher perception and practice	2
Potter et al. (2026)	United States	High school biology teachers (16)	Biology	Mixed methods, participatory design	Generative AI for lesson planning	Teacher perception and practice	2
Pun et al. (2026)	Hong Kong	Junior secondary students, grades 8–9 (68)	Environmental science (climate change)	Mixed methods	GPT-4 with DALL·E	AI-assisted student performance	1, 3
Rappa et al. (2026)	Australia	Secondary students with disability (3)	General science	Qualitative case study	ChatGPT	AI-assisted student performance	2
Tang & Putra (2026)	Indonesia	High school students aged 16–17 (21)	Physics, chemistry, biology	Mixed-methods case study	Custom chatbot (Dialogic Science Teacher)	AI-assisted student performance	3
Tong et al. (2024)	China	Concept-inventory and physics reasoning problems, benchmarked against student data	Physics	AI evaluation	ChatGPT-4	AI capability	2
Tong et al. (2025)	China	Grade 10 students (120)	Physics	Quasi-experimental	ChatGPT-4o	AI-mediated learning outcomes	1, 3
Tschisgale et al. (2025)	Germany	Physics Olympiad problems for secondary students (6)	Physics	AI evaluation	GPT-4o, o1-preview	AI capability	2
Valeri et al. (2025)	Sweden	Upper-secondary students (453 surveyed, 12 interviewed)	Biology, chemistry, physics	Mixed methods	ChatGPT	Student perception	1, 2
Xu, Liu, et al. (2025)	China	Upper-secondary students' physics essay responses (80)	Physics	AI evaluation	ChatGPT-4	AI capability	2

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Journal of Education and Learning Reviews, 3 (6): November-December 2026. ISSN: 3057-0387
Website: https://so19.tci-thaijo.org/index.php/JELS/about
Article ID: e3979

Study	Country	Participants or materials (n)	Discipline	Design	Generative AI tool	Evidence type	RQ
Xu, Reiss, & Lodge (2025)	China	Grade 10 students (189)	General science (scientific creativity)	Instrument development	ChatGPT (as a scoring aid)	AI capability	2

Effects on Conceptual Understanding and Engagement (RQ1)

Seven studies contributed to RQ1, but only two measured learning against a comparison condition, and neither compared generative AI with instruction that used no technology (Table 4). In the first, Ng et al. (2024) compared two intact Secondary 4 classes in Hong Kong over a three-week unit on force and motion. Students who used SRLbot, a chatbot built on ChatGPT to support self-regulated learning, improved on a science knowledge test ($r = .37$, $p = .03$), and between-group comparisons favored SRLbot over a rule-based chatbot on self-regulated learning and motivation measures ($r = .35$ to $.54$, $p < .05$). In the second, Tong et al. (2025) compared two grade 10 classes in China solving scientific knowledge and scientific thinking problems either with ChatGPT-4o or with a peer. Both groups improved, but the gain was smaller with ChatGPT-4o ($d = 0.053$) than with peer collaboration ($d = 0.378$). The two controlled comparisons therefore point in different directions: generative AI outperformed a less capable chatbot but underperformed human collaboration.

The remaining five studies addressed engagement and perceptions rather than learning. Bitzenbauer (2023) found that, after two structured ChatGPT activities, 53 grade 12 physics students agreed more strongly with all but one ChatGPT-related statement, including statements about its benefits; the one-group design and descriptive analysis do not allow conclusions about conceptual understanding, which was not measured. Fekets (2025) reported that 37 Taiwanese upper-secondary students rated a custom physics chatbot as meeting usability expectations on the BUS-15 scale (Cronbach's $\alpha = .709$), with accessibility emerging as the main determinant of a positive experience and slow responses and weak conversational coherence as the main shortcomings. In a survey of 453 Swedish upper-secondary students, Valeri et al. (2025) found that 88.4% had used an AI chatbot for their studies in the previous month and more than 60% used ChatGPT to explain concepts in STEM subjects, most often biology. Fadillah et al. (2024) found strong correlations ($r = .86$ to $.89$) between 334 Indonesian students' perceptions of ChatGPT's convenience and quality, its motivational value, and its accuracy and their self-reported support for higher-order thinking; because all measures were self-reported and cross-sectional, these associations do not establish an effect on understanding. Finally, Pun et al. (2026) examined how 68 Hong Kong junior secondary students read climate-change texts produced by GPT-4 with DALL·E. Students evaluated a cognitive-epistemic text more successfully than a social-institutional one, and many held uninformed ideas about generative AI, equating it with an internet search engine, a picture creator, or a human.

Across RQ1, the direction of evidence was positive for perceptions, usability, and engagement, all measured by self-report, and mixed for learning: one controlled comparison favored generative AI over a rule-based chatbot, and one found smaller gains than peer collaboration. Evidence on conceptual understanding specifically rested on these two quasi-experimental studies of intact classes, both in physics.

Table 4 *Studies Addressing RQ1: Conceptual Understanding and Engagement*

Study	Design (n)	Comparison	Outcome measure	Direction	Principal finding
Ng et al. (2024)	Quasi-experimental, intact classes (74)	Rule-based chatbot	Science knowledge test; self-regulated learning and motivation questionnaire	Positive	Knowledge gain in the SRLbot group ($r = .37$, $p = .03$); between-group differences favored SRLbot ($r = .35$–$.54$)

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Study	Design (n)	Comparison	Outcome measure	Direction	Principal finding
Tong et al. (2025)	Quasi-experimental, intact classes (120)	Peer collaboration	Knowledge (WME) and scientific thinking (iSTAR) tests	Negative relative to comparison	Both groups improved; smaller gain with ChatGPT-4o (d = 0.053) than with peers (d = 0.378)
Bitzenbauer (2023)	One-group pretest–posttest (53)	None	Attitude questionnaire	Positive (descriptive)	Greater agreement with all but one ChatGPT-related statement; conceptual understanding not measured
Fadillah et al. (2024)	Cross-sectional survey (334)	None	Self-reported perceptions and higher-order thinking support	Positive association	r = .86–.89 between perceived ChatGPT qualities and self-reported higher-order thinking support
Fekets (2025)	Mixed-methods pilot (37)	None	BUS-15 usability scale (α = .709)	Positive (descriptive)	Usability expectations met; slow responses and weak coherence noted
Valeri et al. (2025)	Mixed methods (453 surveyed; 12 interviewed)	None	Survey of use and AI knowledge; interviews	Descriptive	88.4% used an AI chatbot in the past month; more than 60% used ChatGPT to explain STEM concepts
Pun et al. (2026)	Mixed methods (68)	Two AI-generated texts	Reading of AI-generated image-texts	Mixed	Better evaluation of the cognitive-epistemic text; uninformed conceptions of generative AI

Pedagogical Strategies, Challenges, and the Reliability of Generative AI (RQ2)

Fourteen studies contributed to RQ2. Their findings fall into three groups: teachers' strategies and perceptions, implementation challenges arising on the student side, and evidence on how reliably generative AI performs the instructional and assessment tasks teachers might delegate to it (Table 5).

Four studies described how teachers integrated generative AI or how they viewed it. Adeyele and Ramnarain (2024) found that some of the 15 Nigerian science teachers they studied had already integrated ChatGPT into the 5E inquiry model and reported benefits, while others were at earlier stages of awareness and adoption. Kilde-Westberg et al. (2025) observed Swedish upper-secondary students consulting ChatGPT during a physics laboratory session on acoustic levitation and found that, although the tool could handle some of the questions and problems that arose, the teacher remained crucial in identifying students' needs and in helping them understand the tool's potential and limitations. Nguyen et al. (2025) used participatory, value-sensitive design with high school students, teachers, and environmental educators in the United States to develop chatbots for an environmental science curriculum, identifying identity development, well-being, human–chatbot relationships, and environmental sustainability as the values that should guide their design. Potter et al. (2026) found that 16 high school biology teachers regarded generative AI as a promising aid for lesson planning but were concerned about its accuracy and ethics, which led to the addition of AI-literacy modules to their professional development. Together with the purpose-built chatbots in Ng et al. (2024) and Tang and Putra (2026), these studies show strategies that ranged from embedding a general-purpose tool in an existing inquiry cycle to designing chatbots around a specific pedagogical aim.

Three studies identified challenges that arose on the student side. Rappa et al. (2026) found that three Australian students with disability made clear and meaningful choices to use ChatGPT in a science class, but metacognitive and cognitive constraints meant that these choices did not always match what they were able to carry out. Valeri et al. (2025) found widespread use of ChatGPT alongside limited knowledge of how it works, although some students had developed prompting strategies to deal with inaccurate or fabricated output. Fekets (2025) reported slow responses and weak conversational coherence in a custom chatbot.

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Journal of Education and Learning Reviews, 3 (6): November-December 2026. ISSN: 3057-0387
Website: https://so19.tci-thaijo.org/index.php/JELS/about
Article ID: e3979

Table 5 *Studies Addressing RQ2: Strategies, Challenges, and Reliability of Generative AI*

Study	Design (n)	Principal finding
Teacher strategies and perceptions		
Adeyele & Ramnarain (2024)	Qualitative (15 teachers)	ChatGPT integrated into the 5E inquiry model at varying stages of adoption
Kilde-Westberg et al. (2025)	Qualitative case study (7 groups)	ChatGPT handled some laboratory questions; teacher remained crucial in identifying needs and limitations
Nguyen et al. (2025)	Qualitative, participatory design (14)	Identity, well-being, human–chatbot relationships, and sustainability identified as design values
Potter et al. (2026)	Mixed methods (16 teachers)	Generative AI seen as a planning aid; concerns about accuracy and ethics
Challenges on the student side		
Rappa et al. (2026)	Qualitative case study (3)	Mismatch between students' choices to use ChatGPT and their metacognitive capability
Valeri et al. (2025)	Mixed methods (453; 12)	Widespread use with limited knowledge of how the technology works
Fekets (2025)	Mixed-methods pilot (37)	Slow responses and weak conversational coherence
Reliability of generative AI for instruction and assessment		
Jere (2025)	AI evaluation (1 exam)	47%–65% on a grade 12 chemistry exam, above the 46% candidate average
Fernández et al. (2024)	AI evaluation (8 exams)	Outperformed students but about 20% of responses contained errors; grading correlation with humans r = .46–.70 vs. .85 between humans
Tong et al. (2024)	AI evaluation (problem sets)	100% on 16 concept-inventory items; 70%–90% on physics reasoning problems
Tschisgale et al. (2025)	AI evaluation (6 problems)	GPT-4o and o1-preview outperformed Physics Olympiad participants on average
Xu, Liu, et al. (2025)	AI evaluation (80 essays)	More consistent but less accurate than human graders against expert scores
Jho & Ha (2024)	AI evaluation (180 arguments)	Higher accuracy for claims and data than for rebuttals, backing, and warrants
Xu, Reiss, & Lodge (2025)	Instrument development (189)	ChatGPT used to support objective scoring of scientific creativity

Seven studies evaluated generative AI's capability on high school science tasks. On examinations and problem solving, all reported performance at or above the level of students. On the November 2023 South African grade 12 chemistry paper, ChatGPT Plus scored 47%, Gemini 51%, and Claude 3.5 Sonnet 65%, against a candidate average of 46% (Jere, 2025). ChatGPT-4 outperformed students' mean grades on eight Spanish university entrance chemistry examinations, yet about 20% of its responses contained errors (Fernández et al., 2024). ChatGPT-4 answered all 16 text-only concept-inventory items correctly and scored 90% and 70% on middle and high school physics reasoning problems, far above benchmark student averages of about 50%, 15%, and 26%, respectively (Tong et al., 2024). Both GPT-4o and o1-preview outperformed German Physics Olympiad participants on average, with o1-preview almost consistently outperforming both GPT-4o and the human benchmark (Tschisgale et al., 2025). Evidence on scoring was less favorable. ChatGPT-4's grades for university entrance chemistry answers correlated with those of two human graders at r = .46 to .49 without a rubric and at about .70 with worked examples, compared with r = .85 between the two human graders themselves (Fernández et al., 2024). In scoring 80 upper-secondary students' physics essays, ChatGPT-

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Journal of Education and Learning Reviews, 3 (6): November-December 2026. ISSN: 3057-0387
Website: https://so19.tci-thaijo.org/index.php/JELS/about
Article ID: e3979

4 was more consistent across repeated scorings than 30 pre-service teachers but less accurate against expert scores in most instances (Xu, Liu, et al., 2025). ChatGPT-3.5 identified claims, data, and qualifiers in 180 upper-secondary students' arguments more accurately than rebuttals, backing, and warrants, the elements that argumentation instruction most seeks to develop (Jho & Ha, 2024). Xu, Reiss, and Lodge (2025) incorporated ChatGPT into the scoring of a new scientific creativity instrument for grade 10 students to improve the objectivity of scoring. These studies concern what generative AI can do rather than what students learn; their relevance to RQ2 lies in what they imply for teachers' assessment practices, namely that unsupervised tasks are open to AI completion and that AI-assisted grading is not yet reliable enough to replace human judgment.

Influence on Critical Thinking and Scientific Inquiry (RQ3)

Five studies contributed to RQ3, and none measured the development of critical thinking or inquiry skills over time with a validated instrument and a comparison group that did not use AI (Table 6). The evidence instead describes reasoning during AI use, reasoning relative to human collaboration, or self-reported thinking.

The most positive findings came from a chatbot designed for dialogue. Tang and Putra (2026) analyzed the chat logs of 21 Indonesian students aged 16 to 17 who interacted with Dialogic Science Teacher, a customized chatbot designed to act as a dialogic partner rather than an answer provider, across physics, chemistry, and biology topics. Perspective-taking, reasoning, arguing, and creative thinking appeared in high proportions of interactions in both participating schools, and qualitative analysis showed students reflecting critically and generating alternative ideas. The study had no comparison condition and examined reasoning during the interaction rather than change over time. Kilde-Westberg et al. (2025) found that ChatGPT could support students' problem solving in a physics laboratory, but only where the teacher helped students recognize its limitations.

Two studies point to the conditions under which reasoning was weaker. In Tong et al.'s (2025) comparison, students working with ChatGPT-4o for the first time tended to use it to obtain answers rather than to explore problems collaboratively; most pasted in screenshots or the question text rather than discussing specific points of confusion, a pattern the authors linked to the smaller gains relative to peer collaboration. Pun et al. (2026) found that junior secondary students' uninformed conceptions of generative AI complicated their evaluation of AI-generated scientific texts. Fadillah et al.'s (2024) correlations between students' perceptions of ChatGPT and their self-reported higher-order thinking support indicate that students believe the tool helps them think, but the measures do not assess critical thinking itself.

Across RQ3, reasoning was richer when the chatbot was designed to prompt dialogue or when a teacher guided its use, and poorer when students used a general-purpose tool to obtain answers. Because the supporting studies were small, mostly uncontrolled, and examined reasoning during or immediately after AI use, this pattern describes the conditions associated with more or less productive reasoning rather than an established effect on skill development.

Table 6 *Studies Addressing RQ3: Critical Thinking and Scientific Inquiry*

Study	Construct	Design (n)	Comparison	Direction	Principal finding
Tang & Putra (2026)	Reasoning and argumentation during chatbot dialogue	Mixed-methods case study (21)	None	Positive (no comparison)	High proportions of perspective-taking, reasoning, arguing, and creative thinking in chat logs
Kilde-Westberg et al. (2025)	Problem solving in the laboratory	Qualitative case study (7 groups)	None	Conditional	Supported problem solving where the teacher guided use
Tong et al. (2025)	Scientific thinking (iSTAR) and problem solving	Quasi-experimental (120)	Peer collaboration	Negative relative to comparison	Smaller gains with ChatGPT-4o; students used it to obtain answers
Pun et al. (2026)	Epistemic evaluation of AI-generated texts	Mixed methods (68)	Two AI-generated texts	Mixed	Uninformed conceptions of generative AI complicated evaluation

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Study	Construct	Design (n)	Comparison	Direction	Principal finding
Fadillah et al. (2024)	Self-reported higher-order thinking support	Cross-sectional survey (334)	None	Positive association (self-report)	Strong correlations with perceived ChatGPT qualities; thinking skills not directly measured

Discussion

This review synthesized 20 empirical studies on generative AI in high school science education. Taken together, the findings do not support a simple conclusion that generative AI either benefits or harms science learning. Positive findings were reported more often in studies in which generative AI was embedded within explicitly defined pedagogical structures, and less favorable findings where students used a general-purpose tool without such structure. This section interprets that pattern through the review's analytical framework, examines the mechanisms and sources of variation that may explain it, compares the findings with earlier reviews, and considers their practical implications. Because the evidence comes from 20 heterogeneous studies, most of them small and non-experimental, the pattern is treated as a recurrent association across studies rather than a demonstrated causal effect.

Interpreting the Findings Through TPACK

The TPACK framework holds that effective technology integration depends on how teachers combine technological, pedagogical, and content knowledge (Mishra & Koehler, 2006), and the reviewed studies indicate where each of these forms of knowledge is called on when generative AI enters a science classroom. Technological content knowledge is required because generative AI is fallible in science content: even strong models in the capability studies made errors in about a fifth of examination responses (Fernández et al., 2024), handled visually presented problems less well (Tong et al., 2024), and scored students' written work less accurately than human graders (Xu, Liu, et al., 2025). The concerns about accuracy raised by the teachers in Potter et al. (2026) reflect this demand. Technological pedagogical knowledge is required because tools of the same general kind supported reasoning when interaction was designed around a pedagogical aim, as in the self-regulated learning chatbot of Ng et al. (2024) and the dialogic chatbot of Tang and Putra (2026), but encouraged answer-seeking when students used one freely (Tong et al., 2025). The varying stages of adoption reported by Adeyele and Ramnarain (2024) suggest that teachers develop this integrated knowledge gradually rather than acquiring it with access to the tool.

This interpretation has an important limitation. In the studies with the most positive findings, the pedagogical structure was built into a custom tool designed by researchers, so the effect of structure cannot be separated from the effects of tool design, novelty, and researcher involvement. No included study compared structured and unstructured use of the same tool with comparable students. TPACK therefore offers a plausible explanation of the pattern rather than a tested one.

Why Structured and Unstructured Use Diverged

The student-level perspectives in the analytical framework help explain why structure appeared to matter. From the perspective of cognitive offloading (Risko & Gilbert, 2016), the clearest contrast in the review is Tong et al.'s (2025) finding that students working with ChatGPT-4o gained less than students working with a peer. Most students pasted problems into the chatbot rather than discussing the points they found difficult, so the tool performed much of the reasoning the task was meant to elicit. A peer, by contrast, cannot simply supply a correct solution, which keeps both partners explaining and disputing ideas. The students were also using an AI chatbot for the first time, which suggests that unfamiliarity and the absence of guidance on how to work with the tool contributed to the result. Similar dynamics have been observed outside science: in a field experiment with high school mathematics students, unrestricted access to GPT-4 improved practice performance but reduced later performance once access was removed, whereas a version with learning safeguards largely avoided this harm (Bastani et al., 2025).

The dialogic chatbot studied by Tang and Putra (2026) illustrates the opposite design choice. It was built to ask questions and present alternative perspectives rather than give answers, so it returned the cognitive work to the student, and students' chat logs showed high levels of reasoning and argumentation. From the perspective of self-regulated learning (Zimmerman, 2002), the chatbot in Ng

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

et al. (2024) supported the planning and monitoring phases of learning, and its advantages were clearest on self-regulation and motivation. Self-regulated use, however, presupposes metacognitive capacity. The students with disability studied by Rappa et al. (2026) chose to use ChatGPT purposefully but could not always carry out what they intended, and many of the younger students in Pun et al. (2026) misunderstood what generative AI is, which complicated their evaluation of its output. Taken together, these studies suggest that generative AI supports reasoning when the design of the tool or of the task keeps students responsible for the thinking, and when students understand the tool well enough to question it. A provisional model summarizing these conditions is presented in the Knowledge Contribution section.

Sources of Variation Across Studies

The pattern also needs to be read against substantial heterogeneity. The included studies spanned the lower and upper secondary years, 12 education systems, and tasks ranging from structured examination questions to open laboratory inquiry. The models studied ranged from GPT-3.5 to reasoning-optimized models such as o1-preview, and capability findings are tied to the version tested: a laboratory study conducted with GPT-3.5 (Kilde-Westberg et al., 2025) and an evaluation of o1-preview on olympiad problems (Tschisgale et al., 2025) describe substantially different technologies. Access also differed, from free versions to paid subscriptions, a difference that Wattanakasiwich et al. (2025) linked to how far teachers engaged with prompt design. Interventions lasted from a single session to five weeks, and outcomes were measured by knowledge tests in only two studies and by self-report, usability ratings, or chat-log analysis in most others. Physics accounted for nearly half the sample, so the findings may not transfer to biology, chemistry, or earth science, where tasks and representations differ. These factors may interact with pedagogical structuring rather than operate independently of it, and the present evidence is too sparse to separate them.

Teacher Mediation and the Limits of AI Capability

A further pattern concerns the teacher. In the two classroom studies that examined it directly, the teacher's role in identifying students' needs and helping them recognize the tool's limitations was described as essential (Kilde-Westberg et al., 2025; Rappa et al., 2026), and teachers in other studies saw the value of generative AI as depending on checks for accuracy and on careful integration (Adeyele & Ramnarain, 2024; Potter et al., 2026). This is a consistent qualitative observation rather than a tested effect: no study compared teacher-guided with unguided use of the same tool, and the observation rests on small samples.

The capability studies add a different kind of argument. Generative AI now matches or exceeds students on many structured science tasks, from a national chemistry examination (Jere, 2025) to olympiad physics problems (Tschisgale et al., 2025), yet it remains less reliable than human judgment in grading (Fernández et al., 2024; Xu, Liu, et al., 2025) and in identifying the more demanding elements of a scientific argument (Jho & Ha, 2024). These findings do not show that teachers make AI use more effective. They do show that capability on its own says little about learning, that unsupervised problem-solving tasks can no longer be assumed to reflect students' own work, and that assessment decisions involving generative AI still require human judgment. As capability increases, the question for teachers shifts from whether the tool can solve a problem to whether its use leaves students doing the thinking the task is meant to develop.

Comparison With Previous Reviews

The findings are consistent with, and add specificity to, earlier reviews. Fock and Siller (2025) found that research on generative AI in secondary STEM education centered on the systems themselves; in the present sample, seven of the 20 studies evaluated AI capability and only two measured learning against a comparison condition, indicating that this imbalance extends to high school science. Cheung et al. (2025) found that research in science education prioritized content knowledge over disciplinary practices, and the present review likewise found the thinnest evidence for critical thinking and inquiry. Pérez-Montesdeoca et al. (2026) concluded that gains in secondary education depend on pedagogically grounded integration, which accords with the pattern reported here. The present findings are more cautious than those of Gunsaldi et al. (2025), who reported that generative AI contributed significantly to achievement and conceptual understanding in middle school science. The difference may reflect this

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

review's separation of capability and perception evidence from learning evidence, the stronger comparison condition in one controlled study (Tong et al., 2025), or differences between age groups.

The contrast with higher education is also instructive. A review in which 69.7% of studies came from higher education found positive effects of ChatGPT most often for knowledge construction and skill development (Lee et al., 2025), and a meta-analysis of 69 experimental studies reported positive effects on academic performance and higher-order thinking (Deng et al., 2025). The high school science evidence reviewed here is thinner and more mixed. One explanation is developmental: adolescents' capacities for self-regulation are still maturing (Steinberg, 2005), which may make answer-seeking use more likely without external structure. Another is methodological: the high school literature contains few controlled comparisons, and the one that compared generative AI with human collaboration found it less effective. Generalizing positive findings from higher education to high school science classrooms therefore appears premature.

Practical Implications

Several implications follow from these findings, although they should be read in light of the limited and heterogeneous evidence on which they rest.

For science teachers, the findings suggest designing generative AI tasks that keep students responsible for reasoning, for example by asking students to evaluate, correct, or argue against AI-generated explanations, as in Bitzenbauer's (2023) activities, or by using tools configured to ask questions rather than supply answers (Tang & Putra, 2026). Open access to a general-purpose chatbot during problem solving appears more likely to encourage answer-seeking (Tong et al., 2025).

For professional development, the findings point to science-specific preparation rather than general orientation to AI tools. The capability and perception studies suggest that teachers need to be able to check AI-generated science content for errors, design student–AI interactions around a learning goal, and redesign assessment tasks, alongside guidance on ethical use, data privacy, and bias (Potter et al., 2026; Tipple et al., 2026). Consistent with TPACK, such preparation is likely to be most useful when it develops technological, pedagogical, and content knowledge together rather than separately.

For assessment, the reviewed evidence supports maintaining human oversight when generative AI is used to grade high-stakes work, particularly where results affect students' progression or grades (Fernández et al., 2024; Xu, Liu, et al., 2025). AI-assisted scoring may be useful for low-stakes feedback or as a first pass, but the appropriate level of oversight depends on the type of task, the stakes involved, and the demonstrated reliability of the system. Because generative AI can now complete many structured problems, assessments intended to capture students' own problem solving may need to be supervised or redesigned (Jere, 2025; Tschisgale et al., 2025).

For curriculum, the gap between students' widespread use of generative AI and their limited understanding of how it works (Pun et al., 2026; Valeri et al., 2025) suggests that AI literacy, including how generative AI produces its output and how to evaluate its scientific claims, should be taught explicitly within science rather than left to informal learning.

For school leaders and education systems, implementation depends on more than individual teachers' initiative. Schools need policies on approved tools, protection of student data, particularly for minors, equitable access across free and paid versions and across languages, academic integrity, and the documentation of AI-assisted assessment, together with periodic review of whether these practices are improving learning (Miao & Holmes, 2023).

Conclusion

This systematic review synthesized 20 empirical studies, published between 2023 and 2026, on generative AI in high school science education. Because the studies were heterogeneous in design, most were small and non-experimental, and none was a randomized controlled trial, its conclusions describe patterns across the available evidence rather than established causal effects.

For RQ1, the evidence on conceptual understanding and engagement was positive for students' perceptions, use, and engagement, which were largely self-reported, but mixed for learning. Of the two studies that measured learning against a comparison condition, one found that a chatbot designed to support self-regulated learning outperformed a rule-based chatbot, and the other found smaller gains

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

with ChatGPT-4o than with peer collaboration. No study compared generative AI with instruction that used no technology.

For RQ2, teachers' strategies ranged from embedding general-purpose tools in inquiry-based lessons and laboratory work to designing chatbots around specific pedagogical aims. The main challenges concerned the accuracy of AI-generated content, the reliability of AI-assisted grading, students' limited understanding of how generative AI works, and teachers' readiness to integrate it. Generative AI performed at or above the level of students on structured examinations and problems but remained less reliable than human judgment in scoring students' work.

For RQ3, the evidence on critical thinking and scientific inquiry was the most limited. Reasoning was richer when a chatbot was designed to prompt dialogue or when a teacher guided its use, and weaker when students used a general-purpose tool to obtain answers. No study measured the development of these skills over time against a comparison group that did not use AI.

Across the three questions, the evidence points more consistently to how generative AI is used than to which tool is used. Pedagogical structuring emerged as a recurrent condition associated with more productive outcomes and may be more consequential than the sophistication of the tool alone. Confirming this pattern will require controlled comparisons of structured and unstructured uses of the same tool, longer interventions, validated measures of reasoning, and studies beyond physics. Until such evidence is available, the central question for high school science education is not whether students will use generative AI, but how its use can be designed so that students, rather than the tool, do the scientific thinking.

Limitations

Several limitations should be considered when interpreting the findings of this review. The first concerns the search. Only three databases were searched, results were limited to English-language, peer-reviewed journal articles, and no citation, reference-list, or grey-literature searching was conducted. Relevant studies indexed elsewhere, published in other languages, or reported in conference proceedings or theses may therefore have been missed, and the small number of included studies partly reflects these restrictions as well as the early stage of the field.

The second concerns access. In Scopus, results were limited to open-access publications, and reports that could not be obtained through open-access sources or institutional access were excluded as not retrieved (Figure 2). Both restrictions select studies on the basis of availability rather than relevance and may introduce availability bias. The direction of any such bias is unknown, but if subscription journals publish systematically different designs or findings, the evidence reviewed here may not represent the full literature.

The third concerns the review process. The protocol was not registered in advance, so the eligibility rules and analytical framework cannot be shown to have been fully specified before screening. An automated keyword-matching procedure and a generative AI assistant were used to organize records before screening; although all eligibility decisions were made independently by two reviewers without access to these provisional classifications, readers should be aware of this use of automation. In addition, synthesis by direction of effect describes how findings are distributed but does not weigh their magnitude or precision.

The fourth concerns the evidence itself. Most included studies were small, short, and non-experimental; only two measured learning against a comparison condition, none was a randomized controlled trial, and many outcomes were self-reported. Nearly half the studies concerned physics, and no single education system contributed more than four studies, which limits transferability to other sciences and settings. Finally, generative AI is changing rapidly. The included studies used models ranging from GPT-3.5 to o1-preview, and findings about AI capability in particular may date quickly as newer models are released.

Directions for Future Research

The limitations of the included studies indicate where stronger evidence is most needed. The most pressing need concerns conceptual understanding and scientific reasoning, the outcomes for which this review found the least controlled evidence. Only two studies compared learning against another condition, and none followed the development of students' critical thinking or inquiry skills over time. Future studies should therefore use randomized or well-matched quasi-experimental designs across several schools, include a comparison condition without generative AI where this is ethically and

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

practically feasible, and assess outcomes with validated tests of conceptual understanding and scientific reasoning rather than self-report alone.

Second, because the pattern identified in this review rests on comparisons between different tools and studies, the design variable itself needs to be tested directly. Studies that compare structured and unstructured uses of the same tool, for example a chatbot that withholds answers and prompts explanation against the same chatbot used freely, would show whether pedagogical structuring accounts for the differences observed here or whether tool design, novelty, or researcher involvement does. The provisional model presented in the Knowledge Contribution section offers testable propositions for such work.

Third, longitudinal studies are needed to examine whether learning gains persist, whether students' use of generative AI changes as their AI literacy develops, and whether reliance on the tool reduces independent problem solving once it is removed, as has been observed in mathematics (Bastani et al., 2025).

Fourth, the evidence needs to extend beyond physics. Chemistry, biology, and earth science were represented by only a handful of studies, although they involve different representations, laboratory practices, and forms of argument. Research in these subjects, and in a wider range of countries and languages, would show whether the patterns found here hold across science education.

Finally, research should examine teachers and equity more directly. Intervention studies of professional development could test whether developing teachers' technological, pedagogical, and content knowledge changes how they structure generative AI use and, in turn, what students learn. Studies of students with disability, students learning science in a second language, and schools with limited access to paid tools would clarify who benefits from generative AI and who may be left behind. Across all of these areas, reporting the model version, prompts, settings, and duration of use would make findings easier to compare and replicate as the technology changes.

Knowledge Contribution

Figure 3 presents the provisional model of conditions associated with more and less productive generative AI use in high school science education that emerges from this review. The model is drawn from the synthesis as a whole rather than from any single study, and it is offered as a set of propositions for empirical testing rather than an established causal account. Its starting point is a single situation: generative AI used in a science lesson or task. The model proposes that the tool alone does not determine the outcome. Instead, outcomes depend on the degree of pedagogical structuring of student–AI interaction, meaning the extent to which the tool, the task, or the teacher keeps students responsible for the thinking.

The model identifies four groups of moderating conditions that may influence how much structuring takes place: teachers' technological, pedagogical, and content knowledge (TPACK); students' AI literacy, including the metacognitive capacity to question AI output; the design of tasks and the stakes of assessment; and access and school support, including paid or free versions of tools, language support, and school policy. These conditions are drawn from patterns across the included studies and the background literature, and their relative importance has not been established.

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Journal of Education and Learning Reviews, 3 (6): November-December 2026. ISSN: 3057-0387
Website: https://so19.tci-thaijo.org/index.php/JELS/about
Article ID: e3979

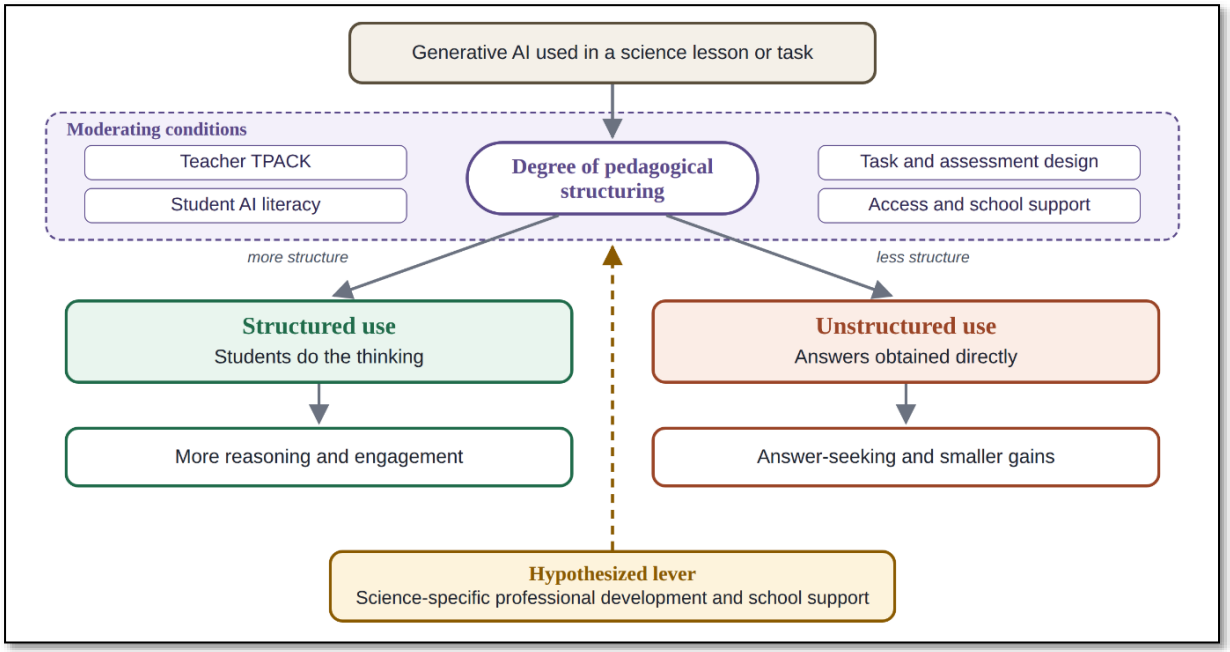


Figure 3. Provisional model of conditions associated with more and less productive generative AI use in high school science education

From this point, the model branches into two pathways. In structured use, students do the thinking, for example through a chatbot designed to prompt dialogue (Tang & Putra, 2026), support for planning and monitoring learning (Ng et al., 2024), or teacher guidance in recognizing the tool's limitations (Kilde-Westberg et al., 2025); these studies reported richer reasoning, argumentation, and engagement. In unstructured use, students have open access to a general-purpose tool and can obtain answers directly. In the clearest example in the review, this was associated with answer-seeking and smaller learning gains than peer collaboration (Tong et al., 2025), a pattern consistent with cognitive offloading. Both pathways are shown with solid arrows because they reflect associations reported across the included studies.

Finally, the model identifies science-specific professional development and school support as a hypothesized lever for practice. If structured use depends on teachers' integrated knowledge and students' AI literacy, then professional development and school support that build these capacities should make structured use more likely. No included study tested this proposition, which is why it appears as a dashed arrow in Figure 3. Testing it, together with direct comparisons of structured and unstructured use of the same tool, would show whether the model holds.

Recommendation

Based on the findings and discussion presented in this review, the following recommendations are offered.

1. Science teachers should design generative AI use around a clear learning goal rather than give students open access to a general-purpose chatbot. Tasks that ask students to explain, evaluate, correct, or argue against AI-generated responses keep students responsible for the thinking.
2. Schools and education authorities should provide science-specific professional development rather than general orientation to AI tools. Such training should help teachers check AI-generated science content for errors, design tasks that require students to reason with the tool, redesign assessment,

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

and teach students how generative AI works. These programmes should be evaluated as they are introduced, since their effects have not yet been tested.

3. Assessment practices should keep human judgment at the center. Generative AI may support low-stakes feedback, but grades that affect students' progression should be reviewed by a teacher. Unsupervised tasks meant to capture students' own problem solving may need to be redesigned or supervised.

4. Curriculum developers should teach AI literacy explicitly within science, including how generative AI produces its output, why it can be wrong, and how to check its claims against evidence.

5. School leaders should establish clear policies on approved tools, student data protection, equitable access, and academic integrity, and review periodically whether these practices are improving learning.

6. Researchers should test the provisional model directly. Priorities include comparing structured and unstructured use of the same tool, using larger multi-site designs with validated measures of learning and reasoning, testing whether professional development leads to more structured use, and extending research beyond physics.

References

Adeyele, V. O., & Ramnarain, U. (2024). Exploring the integration of ChatGPT in inquiry-based learning: Teacher perspectives. *International Journal of Technology in Education, 7*(2), 200–217. https://doi.org/10.46328/ijte.638

Ajose, S. T., Apata, O. E., Saidu, G. O., & Orobator, E. S. (2026). Exploring teachers' awareness and perceptions of ChatGPT in K–12 STEM education. *International Journal of Technology in Education, 9*(1), 260–278. https://doi.org/10.46328/ijte.5077

Azizi, M. N., Faizi, A., & Sari, U. (2026). Artificial intelligence in physics education (2015–2025): Systematic review of trends, applications, and challenges. *Journal of Education in Science, Environment and Health, 12*(2), 174–197. https://doi.org/10.55549/jeseh.905

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

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

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ, 368*, Article l6890. https://doi.org/10.1136/bmj.l6890

Cheung, K. K. C., Zerouali, A., Koenen, J., & Erduran, S. (2025). Do generative artificial intelligence (GenAI) and science education mix? A systematic review of the literature. *Studies in Science Education*. Advance online publication. https://doi.org/10.1080/03057267.2025.2578091

Co, S. J., Dabu, J. P., & Sanchez, P. J. (2026). AI literacy and academic performance: A cross-sectional analysis of senior high school students. *International Journal of Technology in Education, 9*(2), 476–490. https://doi.org/10.46328/ijte.5335

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, Article 105224. https://doi.org/10.1016/j.compedu.2024.105224

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

Fekets, G. (2025). Development of an artificial intelligence-supported chatbot as an interactive learning platform in STEM education: Exploring usability and student experience. *Journal of Baltic Science Education, 24*(5), 878–893. https://doi.org/10.33225/jbse/25.24.878

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Fernández, A. A., López-Torres, M., Fernández, J. J., & Vázquez-García, D. (2024). ChatGPT as an instructor's assistant for generating and scoring exams. *Journal of Chemical Education, 101*(9), 3780–3788. https://doi.org/10.1021/acs.jchemed.4c00231

Fock, A., & Siller, H.-S. (2025). Generative artificial intelligence in secondary STEM education in the light of human flourishing: A scoping literature review. *International Journal of STEM Education, 12*, Article 67. https://doi.org/10.1186/s40594-025-00589-5

Fütterer, T., Bardach, L., Kuhn, J., Keller, S. D., & Gerjets, P. (2026). Enhancing school students' self-regulated learning through generative AI support: A randomized controlled trial. *Educational Psychology Review, 38*(1), Article 42. https://doi.org/10.1007/s10648-026-10133-8

Garcia, R., Fojas, J. G., Formanes, J. N., Garcia, D., Habol, C., & Orapa, L. (2025). Perspective of science educators and students on the use of ChatGPT: A study of artificial intelligence. *International Journal of Technology in Education and Science, 9*(4), 474–491. https://doi.org/10.46328/ijtes.649

Gunsaldi, M. S., Guner, E. G., Uckan, M., & Bati, K. (2025). The impact of generative AI applications on student learning outcomes in science education: A systematic review. *Journal of Education in Science, Environment and Health, 11*(3), 196–208. https://doi.org/10.55549/jeseh.840

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). *Mixed Methods Appraisal Tool (MMAT), version 2018: User guide*. McGill University. https://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/127916259/MMAT_2018_criteria-manual_2018-08-01_ENG.pdf

Jang, H., & Choi, H. (2025). A double-edged sword: Physics educators' perspectives on utilizing ChatGPT and its future in classrooms. *Journal of Science Education and Technology, 34*(2), 267–283. https://doi.org/10.1007/s10956-024-10173-1

Jere, S. (2025). Evaluating artificial intelligence large language models' performances in a South African high school chemistry exam. *Eurasia Journal of Mathematics, Science and Technology Education, 21*(2), Article em2582. https://doi.org/10.29333/ejmste/15932

Jho, H., & Ha, M. (2024). Towards effective argumentation: Design and implementation of a generative AI-based evaluation and feedback system. *Journal of Baltic Science Education, 23*(2), 280–291. https://doi.org/10.33225/jbse/24.23.280

Kilde-Westberg, S., Johansson, A., & Enger, J. (2025). Generative AI as a lab partner: A case study. *Physical Review Physics Education Research, 21*(2), Article 020119. https://doi.org/10.1103/ggy1-3kjk

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lee, H.-Y., Huang, Y.-M., & Wu, T.-T. (2025). ChatGPT in education: A systematic review of current landscape, limitations and future directions through general system theory lens. *European Journal of Education, 60*(4), Article e70262. https://doi.org/10.1111/ejed.70262

Marzano, D. (2026). Generative artificial intelligence (GAI) in teaching and learning processes at the K–12 level: A systematic review. *Technology, Knowledge and Learning, 31*, 789–829. https://doi.org/10.1007/s10758-025-09853-7

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. *British Journal of Educational Technology, 55*(4), 1328–1353. https://doi.org/10.1111/bjet.13454

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

Nguyen, H., Nguyen, V., Ludovise, S., & Santagata, R. (2025). Value-sensitive design of chatbots in environmental education: Supporting identity, connectedness, well-being and sustainability. *British Journal of Educational Technology, 56*(4), 1370–1390. https://doi.org/10.1111/bjet.13568

Osborne, J. (2014). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education, 25*(2), 177–196. https://doi.org/10.1007/s10972-014-9384-1

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews, 5*, Article 210. https://doi.org/10.1186/s13643-016-0384-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Pérez-Montesdeoca, H., Rodriguez-Rodriguez, D., & Fernández-Sogorb, A. (2026). The integration of generative artificial intelligence in secondary education: A systematic review. *Contemporary Educational Technology, 18*(3), Article ep667. https://doi.org/10.30935/cedtech/18746

Potter, A. H., Arner, T., McCarthy, K. S., & McNamara, D. S. (2026). Integrating reading, writing, and digital tools in science: A participatory-design study of the InSPECT framework. *Education Sciences, 16*(1), Article 6. https://doi.org/10.3390/educsci16010006

Pun, J., Cheung, K. K. C., & Kenneth-Li, W. (2026). Secondary students' reading of socio-scientific image-texts on climate change in a GPT-4 scenario. *Research in Science Education, 56*(1), 183–202. https://doi.org/10.1007/s11165-025-10258-w

Rappa, N. A., Nonis, K. P., Tang, K.-S., Cooper, G., Cooper, M., & Sims, C. (2026). Can generative AI support the learning agency of students with disability? A case study of an Australian secondary school. *British Journal of Educational Technology, 57*(4), 924–942. https://doi.org/10.1111/bjet.70048

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Shankar, S. K., Pothancheri, G., Sasi, D., & Mishra, S. (2025). Bringing teachers in the loop: Exploring perspectives on integrating generative AI in technology-enhanced learning. *International Journal of Artificial Intelligence in Education, 35*(1), 155–180. https://doi.org/10.1007/s40593-024-00428-8

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Steinberg, L. (2005). Cognitive and affective development in adolescence. *Trends in Cognitive Sciences, 9*(2), 69–74. https://doi.org/10.1016/j.tics.2004.12.005

Tang, K.-S., & Putra, G. B. S. (2026). Generative AI as a dialogic partner: Enhancing multiple perspectives, reasoning, and argumentation in science education with customized chatbots. *Journal of Science Education and Technology, 35*(1), 128–140. https://doi.org/10.1007/s10956-025-10240-1

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology, 8*, Article 45. https://doi.org/10.1186/1471-2288-8-45

Tipple, C., Keane, T., Wang, T., & Cerovac, M. (2026). Teachers' perceptions of generative AI in gender-inclusive STEM education: A grounded theory study. *Education and Information Technologies, 31*(8), 2399–2434. https://doi.org/10.1007/s10639-026-13908-2

Tong, D., Jin, B., Tao, Y., Ren, H., Islam, A. Y. M. A., & Bao, L. (2025). Exploring the role of human-AI collaboration in solving scientific problems. *Physical Review Physics Education Research, 21*(1), Article 010149. https://doi.org/10.1103/PhysRevPhysEducRes.21.010149

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979

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

Tschisgale, P., Maus, H., Kieser, F., Kroehs, B., Petersen, S., & Wulff, P. (2025). Evaluating GPT- and reasoning-based large language models on physics olympiad problems: Surpassing human performance and implications for educational assessment. *Physical Review Physics Education Research, 21*(2), Article 020115. https://doi.org/10.1103/6fmx-bsnl

Valeri, F., Nilsson, P., & Cederqvist, A.-M. (2025). Exploring students' experience of ChatGPT in STEM education. *Computers and Education: Artificial Intelligence, 8*, Article 100360. https://doi.org/10.1016/j.caeai.2024.100360

Wattanakasiwich, P., Kaewkhong, K., & Katwibun, D. (2025). Physics instructors' acceptance and implementation of generative AI. *Physical Review Physics Education Research, 21*(1), Article 010155. https://doi.org/10.1103/r2fn-kdy4

Xu, S., Reiss, M. J., & Lodge, W. (2025). Comprehensive Scientific Creativity Assessment (C-SCA): A new approach for measuring scientific creativity in secondary school students. *International Journal of Science and Mathematics Education, 23*(2), 293–319. https://doi.org/10.1007/s10763-024-10469-z

Xu, Y., Liu, L., Xiong, J., & Zhu, G. (2025). Graders of the future: Comparing the consistency and accuracy of GPT4 and pre-service teachers in physics essay question assessments. *Journal of Baltic Science Education, 24*(1), 187–207. https://doi.org/10.33225/jbse/25.24.187

Yusuf, A., Pervin, N., Román-González, M., & Md Noor, N. (2024). Generative AI in education and research: A systematic mapping review. *Review of Education, 12*(2), Article e3489. https://doi.org/10.1002/rev3.3489

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Citation: Aniza C. Cargason, A.C., Dumam-ag, S.F., Amol, J.U., Dirampatun, S.M. & Sabasales, M. (2026). Generative Artificial Intelligence in High School Science Education: A Systematic Review of Learning Outcomes, Pedagogical Strategies, and Critical Thinking Development. Journal of Education and Learning Reviews, 3 (6), e3979; DOI: https://doi.org/10.60027/jelr.2026.e3979