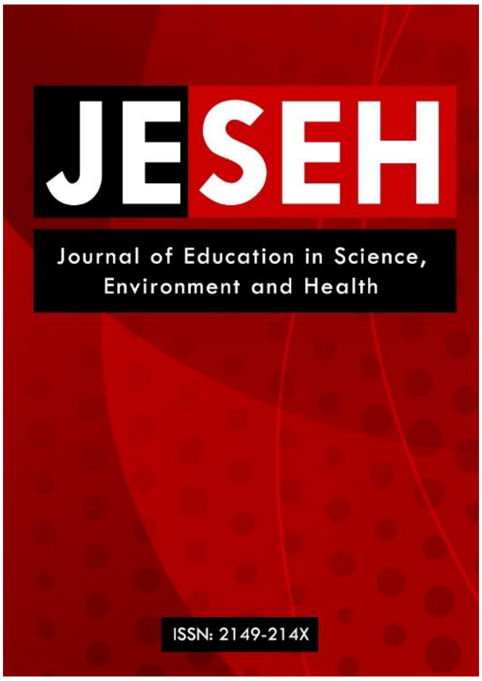

ISSN: 2149-214X

Journal of Education in Science, Environment and Health

www.jeseh.net

How Generative AI Supports Scientific Inquiry and Science Practices in Science Education: A Scoping Review of Empirical Studies

M. Sait Gokalp
Kutahya Dumlupinar University

To cite this article:

Gokalp, M.S. (2026). How generative AI supports scientific inquiry and science practices in science education: A scoping review of empirical studies. *Journal of Education in Science, Environment and Health (JESEH), 12*(3), 265-275. https://doi.org/10.55549/jeseh.912

Journal of Education in Science, Environment and Health
Volume 12, Issue 3, 2026 https://doi.org/10.55549/jeseh.912

How Generative AI Supports Scientific Inquiry and Science Practices in Science Education: A Scoping Review of Empirical Studies

M. Sait Gokalp

Article Info

Article History

Published:
01 July 2026

Received:
04 April 2026

Accepted:
11 June 2026

Keywords

GenAI
Science education
Scoping review

Abstract

Generative artificial intelligence (GenAI) has rapidly entered science education, but its role in supporting scientific inquiry and science practices remains unclear. This scoping review mapped empirical studies on how GenAI has been used to support inquiry- and practice-oriented work in science education. Guided by JBI scoping review principles and PRISMA-ScR, the review used a reproducible open-data search route through OpenAlex and Crossref, supplemented by targeted manual discovery. Records published from 30 November 2022 onward were screened against predefined criteria, resulting in 14 included empirical journal articles. The reviewed studies clustered around three main pedagogical roles: learner-facing inquiry partners, teacher-facing design and co-design tools, and assessment or feedback agents. Investigation planning, explanation, and inquiry scaffolding were more common than data analysis, argumentation, and explicit epistemic reflection. Overall, the literature suggests that GenAI is currently used more often to structure, prompt, or evaluate inquiry than to support deeper disciplinary participation. Productive use depended heavily on teacher mediation, prompt quality, and learners' critical oversight. The review identifies both practical opportunities and persistent epistemic gaps for future research.

Introduction

Generative artificial intelligence (GenAI) has rapidly become a visible part of science education research and practice. Following the public release of ChatGPT on 30 November 2022, researchers increasingly began to examine how GenAI might support science teaching, learning, assessment, feedback, and teacher preparation. This rapid growth has made GenAI one of the most discussed emerging technologies in contemporary science education and has intensified questions about both its pedagogical value and its epistemic implications (Lee et al., 2026).

In science education, the significance of GenAI extends beyond technological novelty. The goals of science education are not limited to transmitting scientific concepts or improving short-term achievement. Science education also aims to engage learners in disciplinary ways of knowing and doing, including asking questions, planning investigations, analyzing data, constructing explanations, using models, evaluating evidence, and participating in argumentation. For this reason, the role of GenAI in science education should not be evaluated only in terms of whether it improves general learning outcomes. It should also be examined in relation to whether it meaningfully supports scientific inquiry and science practices.

This disciplinary lens matters because science education is especially sensitive to how knowledge is produced, justified, and communicated. In many educational contexts, a tool may be judged primarily by whether it improves efficiency, access, or short-term performance. In science education, however, a tool must also be judged by whether it helps learners participate in the practices through which scientific knowledge is developed and critiqued. A response that is fluent or plausible is not necessarily evidence-based, and a lesson that appears inquiry-oriented is not necessarily strong in modeling, explanation, or argumentation. As a result, a review that remains at the level of general AI integration risks overlooking whether GenAI is aligned with the disciplinary core of science learning.

Recent review studies have advanced the field considerably. Broad reviews have already examined the role of artificial intelligence in science teaching and learning (Almasri, 2024), and more recent work has synthesized the general use of GenAI in science education (Cheung et al., 2025). Gunsaldi et al. (2025) reviewed the impact of GenAI applications on student learning outcomes in science education, while Duran and Dikmenli (2025) reviewed the use of artificial intelligence in biology education more broadly. Other recent reviews have also highlighted unresolved concerns related to disciplinary epistemic engagement, science identity, and the possibility that AI tools may support interest or content uptake more strongly than participation in scientific practices (Cheung, 2026; Lee et al., 2026).

These developments are important, but they also reveal a gap. Existing reviews tend to emphasize broad integration trends, student outcomes, or AI use across a wide range of educational goals. As a result, less attention has been paid to a more discipline-sensitive question: how GenAI is being used specifically to support scientific inquiry and science practices in science education. This gap matters because inquiry and practice are not peripheral features of science education. They are central to how science is learned, communicated, and justified in classrooms.

The emergence of inquiry-focused empirical work makes this gap timely to address. For example, Chang et al. (2026) showed that GenAI can be designed to provide feedback on students' epistemic understanding in scientific inquiry. Recent review work suggests that scientific practices, epistemic affect, and practice-oriented engagement remain underrepresented in the broader AI-science-education literature (Cheung, 2026). Recent field-level mapping has likewise pointed to continuing concerns around teacher preparedness, ethical decision making, and the challenge of integrating AI into science education without flattening disciplinary complexity (Lee et al., 2026). Taken together, these developments suggest that the field is now mature enough to benefit from a focused synthesis of how GenAI is positioned within the practice-oriented goals of science education, but still heterogeneous enough to warrant a scoping rather than effect-focused review approach.

The present review is intended to make that contribution. Rather than treating GenAI as a generic educational technology, it maps how GenAI is being embedded in inquiry-oriented tasks, science-practice routines, and pedagogical designs that are distinctive to science education. In doing so, it shifts the analytical focus from simple adoption or outcome claims to a more discipline-sensitive set of questions: what kinds of scientific work GenAI is being asked to support, what roles are assigned to it in that work, and where the emerging empirical literature still falls short.

Accordingly, this scoping review examined the following questions:

1. In what contexts has generative AI been used in science education?
2. Which dimensions of scientific inquiry and science practices are most frequently supported by generative AI?
3. What pedagogical roles are assigned to generative AI in the reviewed studies?
4. What cognitive, epistemic, affective, and assessment-related outcomes are reported?
5. What limitations, risks, and research gaps are highlighted in the literature?

Method

This study adopted a scoping review design to map the emerging empirical literature on the use of generative artificial intelligence in science education. A scoping review was appropriate because the literature has grown rapidly since late 2022, remains conceptually heterogeneous, and includes a wide range of educational purposes, study designs, and science-learning contexts. The review was guided by JBI scoping review principles (Peters et al., 2020; Peters et al., 2021) and structured in line with the PRISMA Extension for Scoping Reviews (PRISMA-ScR) (Tricco et al., 2018).

The review focused on English-language empirical journal articles published from 30 November 2022 onward. This starting point was selected because the public release of ChatGPT marked a major shift in the visibility and educational uptake of generative AI tools. The search was conducted through a reproducible open-data route rather than a subscription-database route. Specifically, records were identified through the OpenAlex works API and the Crossref works API, supplemented by targeted manual discovery through publisher pages and science-education journal searches. It should therefore be understood as an open-data scoping search rather than as a Scopus, Web of Science, and ERIC review.

The search strategy combined three concept groups: generative AI terms, science education terms, and inquiry- or practice-related terms. Across the scripted search process, six query clusters were used: core inquiry, core practices, argumentation and explanation, assessment support, lesson design, and laboratory support. Shared search limits included publication date from 30 November 2022 onward, English language, and journal-article formats where the source API supported such filtering. The scripted search produced 900 raw records, which were deduplicated to 684 candidate records. A high-confidence prescreen then reduced this pool to 58 records for closer inspection. In parallel, 20 additional candidate records were identified through targeted manual discovery. These two routes were later merged and deduplicated at the retained-record stage.

Studies were included if they were peer-reviewed empirical journal articles written in English; situated in science education contexts such as general science, biology, chemistry, physics, environmental science, or science teacher education; examined a generative AI tool, large language model, or AI chatbot; and addressed scientific inquiry, science practices, or closely related disciplinary processes such as explanation, modeling, evidence evaluation, investigation design, or data interpretation. Studies were excluded if they were editorials, conceptual papers, commentaries, conference abstracts, dissertations, or non-peer-reviewed texts. Studies were also excluded if they focused on AI in a broad sense without a clear generative-AI component, if they were not situated in science education, or if they addressed AI literacy without linking the work to science learning, inquiry, or science practices.

Screening proceeded in two stages. First, titles and abstracts were screened against the predetermined criteria. Of the 58 high-confidence open-data records, 40 were excluded and 18 were retained for closer review. Of the 20 manually discovered candidate records, 13 were excluded and 7 were retained. This yielded 25 retained records before cross-source deduplication. After merging these records, 19 unique priority records remained for closer eligibility review. Second, full-text articles were consulted for these 19 records in order to make final inclusion decisions and extract study characteristics. This process resulted in 14 included studies and 5 excluded studies in the final working corpus. Table 1 summarizes the PRISMA-style flow used in this review.

Table 1. PRISMA-style study selection flow summary

Stage	Count	Notes
Raw records identified through OpenAlex and Crossref	900	Scripted open-data retrieval across six query clusters
Deduplicated open-data candidate records	684	Initial deduplication at the metadata level
High-confidence open-data records advanced for screening	58	Prescreened for closer title/abstract review
Additional manually discovered candidate records	20	Identified through targeted journal and publisher discovery
Open-data records excluded at title/abstract screening	40	Removed for lack of science-education, GenAI, or inquiry/practice fit
Open-data records retained for closer review	18	Advanced to eligibility review
Manual candidate records excluded at screening	13	Removed for insufficient inquiry/science-practice centering
Manual candidate records retained for closer review	7	Advanced to eligibility review
Retained records before cross-source deduplication	25	Combined open-data and manual routes
Cross-source duplicates removed	6	Duplicates identified after merging retained sets
Unique full-text or closer-eligibility records assessed	19	Priority corpus for final decision making
Records excluded at closer eligibility review	5	Excluded for insufficient centering on inquiry or science practices
Final included empirical studies	14	Studies included in the thematic synthesis

Data from included studies were charted using a structured extraction form. The charting categories included bibliographic details, science domain, educational level, setting, participant group, GenAI tool, pedagogical role of GenAI, targeted inquiry or science practice, task design, reported outcomes, and reported risks or limitations. Particular attention was given to whether GenAI was positioned as a tutor, co-inquirer, feedback provider, assessment tool, lesson-design assistant, or content-generation aid, and to which inquiry or practice dimensions were most clearly supported. When study fit was uncertain, records were carried forward to closer eligibility review rather than being excluded prematurely.

The extracted data were synthesized using descriptive mapping and qualitative thematic analysis. Descriptive synthesis was used to summarize publication timing, disciplinary spread, educational levels, and types of GenAI tools. Thematic synthesis was then used to identify recurring pedagogical roles, the science practices most often emphasized, reported outcomes, and persistent boundary conditions such as teacher mediation, prompting quality, and epistemic risk. In line with common scoping review practice, the purpose of the review was to map the shape of the literature rather than to conduct a formal quality appraisal for effect estimation.

Results

Overview of Included Studies

Fourteen empirical studies met the inclusion criteria. The publication pattern was strongly concentrated in the most recent years: one study was published in 2023, two in 2024, eight in 2025, and three in early 2026. This distribution suggests that inquiry- and practice-oriented work on GenAI in science education is still a very recent literature, with most of the empirical base emerging after the first wave of broad commentary on ChatGPT in education.

The included studies spanned elementary, middle school, secondary/high-school, teacher-education, and undergraduate contexts. Most were anchored in school-level learning environments, while three were situated in undergraduate settings and several others focused on teacher-facing design or assessment work. This spread is helpful because it shows that GenAI-supported inquiry is not confined to a single age group or science domain. At the same time, it also highlights that the field is still fragmented: the reviewed studies differ substantially in educational level, task design, participant type, and outcome focus. Sample sizes ranged from small exploratory cohorts and interview-based studies to a larger performance dataset of 517 student responses in physics feedback research.

Disciplinarily, physics-related settings formed the clearest cluster, especially in guided inquiry, experiment design, density-concept instruction, and automated feedback. Biology and environmental science were also visible through lesson planning, laboratory assistance, and causal-modeling work. The remaining studies addressed chemistry/computational modeling, socioscientific inquiry, general science teacher integration, science inquiry assessment, and K-12 science-assessment critique. This pattern suggests that GenAI has already entered multiple corners of science education, but not yet in a balanced way across science domains or practice types.

Most studies examined general-purpose LLMs or LLM/chatbot variants, especially ChatGPT and closely related GPT-based tools. Thirteen of the 14 included studies relied on an LLM-family system, while only one study used a rule-based and AI-chatbot hybrid. One study emphasized a multimodal LLM, and one centered an open-source local model for physics feedback. Human guidance was explicit in 13 of the 14 studies, whereas prompt transparency was clearly reported in only 8 studies. This means that the field is simultaneously highly dependent on human orchestration and still inconsistent in documenting how GenAI was actually prompted and pedagogically configured. Table 2 presents the study identifiers used throughout the thematic synthesis.

Pedagogical Roles Assigned to GenAI

Three broad pedagogical role clusters were visible across the corpus. The first cluster consisted of six learner-facing inquiry-partner studies. In these studies, chatbots or LLMs supported guided inquiry, information search and justification, computational modeling, open-ended experiment design, causal modeling, or laboratory questioning (Chang et al., 2023; Stadler et al., 2024; Haraldsrud & Odden, 2025; Min et al., 2025; Nguyen et al., 2025; Doğru & Faulconer, 2026). In this role, GenAI was usually positioned as a scaffold around scientific work rather than as an autonomous replacement for student reasoning. The strongest examples involved bounded participation in specific tasks such as formulating inquiry steps, exploring alternative model structures, or asking laboratory questions during an ongoing activity.

The second cluster consisted of five teacher-facing design or co-design studies. In these studies, teachers or researchers used GenAI to generate, revise, or evaluate inquiry-oriented lesson plans and lesson-study materials (Adeyele & Ramnarain, 2024; Kahraman & Kıyıcı, 2025; Großmann et al., 2025; Cheng, 2025; Kim, 2026). Here the main attraction of GenAI was speed, idea generation, adaptation to curricular goals, and support for planning inquiry-oriented instruction. However, this cluster also showed that inquiry alignment was uneven. AI-generated plans often looked promising at the surface level but still required substantial pedagogical revision to sustain curiosity, coherence, procedural inquiry, or epistemic focus. In other words, design efficiency was easier to obtain than reliable inquiry quality.

The third cluster consisted of three assessment and feedback studies. This cluster included critique of science assessments for NGSS alignment, automated feedback on student physics responses, and scoring with dialogic feedback for scientific inquiry instruments (Nguyen & Hayward, 2025; Meyer et al., 2025; Chang et al., 2026). Compared with some of the broader inquiry uses, this cluster appeared comparatively mature and operational, especially in cases where the task structure and scoring criteria were clearly defined. These studies suggest that GenAI may currently be most defensible in bounded evaluative tasks where outputs can be checked against rubric-like expectations or human ratings.

Table 2. Characteristics of included studies

Study ID	Study	Level	Domain	GenAI tool	Pedagogical role	Main task focus
C002	Chang et al. (2023)	Elementary gifted education	Physics/sound transmission	Inquirybot	Guided-inquiry facilitator	Guided inquiry on sound transmission
C003	Stadler et al. (2024)	Undergraduate	Socioscientific inquiry	ChatGPT 3.5	Information source and reasoning partner	Inquiry on nanoparticles in sunscreen and justified recommendations
C005	Adeyele and Ramnarain (2024)	Secondary education	Science education	ChatGPT	Inquiry-support planning and integration aid	Teacher perspectives on ChatGPT within 5E inquiry-based learning
C007	Nguyen and Hayward (2025)	K-12	Science assessment	GPT-4 Vision or MLLM	Assessment tool and feedback provider	Critique of science assessments for NGSS-aligned sense-making
C008	Meyer et al. (2025)	High school	Physics	OpenChat 3.6; GPT-4o subset	Automated feedback provider	Feedback generation for short written physics responses
C009	Haraldsrud and Odden (2025)	Undergraduate	Chemistry/computational modeling	ChatGPT-4.0	Modeling scaffold and reasoning partner	Computational chemistry modeling task
C012	Kahraman and Kıyıcı (2025)	Middle school	Science	ChatGPT	Lesson-planning assistant	Inquiry-based lesson-plan generation
C013	Min et al. (2025)	Middle school	Physics	Generative AI assistant	Co-inquirer and tutor	Open-ended experiment design
C014	Großmann et al. (2025)	Secondary education	Biology	ChatGPT-4o	Lesson-planning assistant	Inquiry-based biology lesson-plan generation
C015	Cheng (2025)	Grade 7	Density concept	ChatGPT	Lesson-study support and co-designer	Density lesson-study implementation
C016	Kim (2026)	Teacher education	Environmental science agency	Large language model AI	Lesson-design assistant and co-designer	Environmental-science-agency lesson design
C017	Nguyen et al. (2025)	High school	Environmental science	LLM-based chatbots	Causal-modeling support and co-inquiry partner	Climate-change causal modeling
C018	Doğru and Faulconer (2026)	Undergraduate	Biology	ChatGPT	Virtual teaching assistant	Laboratory question answering on enzyme activity
C019	Chang et al. (2026)	Elementary school	Science inquiry	ChatGPT-4o	Assessment and dialogic feedback partner	Scoring and feedback on a scientific inquiry instrument

Taken together, these three clusters show that the pedagogical role of GenAI in science education is not singular. It shifts depending on whether the goal is to help learners move through a scientific task, help teachers prepare inquiry-oriented instruction, or help educators evaluate scientific work. What unifies the clusters is that GenAI is rarely treated as an independent instructional actor. Across roles, it functions more often as a support infrastructure that extends planning, prompting, annotation, or feedback.

Science Practices Supported and Underrepresented

The coding pattern suggested that the literature currently emphasizes some science-practice dimensions much more than others. Investigation planning appeared most often, with 7 of the 14 studies explicitly supporting tasks such as experiment design, lesson planning for inquiry, or planning of science learning sequences. Explanations and information evaluation or communication each appeared in 6 studies, while asking questions and modeling appeared in 4 studies each. In contrast, explicit support for data analysis was identified in only 2 studies, computational or mathematical thinking in only 2 studies, and argumentation in only 1 study. Table 3 shows this imbalance clearly.

This distribution is important because it indicates where GenAI is currently entering inquiry work. The reviewed tools were more often used at the front end of inquiry, where learners or teachers need help generating options, organizing a task, or drafting explanations, than at the back end, where stronger evidence interpretation, justification, and argumentative scrutiny are required. Even when information evaluation or communication was present, it was often embedded in guided prompting or critique tasks rather than in extended disciplinary discourse. As a result, the corpus gives a stronger picture of GenAI as a starter, organizer, or explainer than as a sustained partner in evidence-based scientific reasoning.

The cross-cutting codes point in a similar direction. Inquiry scaffolding was present in 10 studies and teacher mediation in 9 studies, indicating that much of the literature treats GenAI as support infrastructure around inquiry rather than as a stand-alone disciplinary agent. Feedback on practice appeared in only 4 studies, and explicit epistemic-understanding work was rare, appearing clearly in only 2 studies, even though epistemic engagement is central to practice-oriented science learning. Prompt transparency was reported in only 8 studies, which suggests that the reporting quality of GenAI use remains uneven and that the literature is not yet consistently documenting one of the main determinants of model performance.

Table 3. Science-practice and cross-cutting theme frequencies

Code or theme	n	Study IDs	Interpretation
Investigation planning	7	C002, C005, C012, C013, C014, C016, C018	The strongest practice emphasis in the corpus; GenAI is often used to structure inquiry design and lesson planning.
Explanations	6	C002, C003, C005, C008, C015, C017	Explanation support appears frequently, especially in guided inquiry, feedback, and causal reasoning tasks.
Information evaluation and communication	6	C003, C005, C007, C009, C016, C017	GenAI is regularly positioned as a tool for evaluating, critiquing, or communicating science information.
Asking questions	4	C002, C005, C014, C018	Question-focused support appears mainly in inquiry facilitation and laboratory assistance.
Modeling	4	C009, C014, C015, C017	Modeling appears in chemistry, biology lesson planning, lesson study, and environmental causal reasoning.
Data analysis	2	C013, C018	Data-analysis support is present but clearly underrepresented.
Argumentation	1	C003	Explicit argumentation work is rare in the current empirical base.
Inquiry scaffolding	10	C002, C005, C009, C012, C013, C014, C015, C016, C017, C018	The dominant cross-cutting pattern is scaffolding rather than autonomous disciplinary participation.
Teacher mediation	9	C002, C005, C007, C009, C012, C014, C015, C016, C017	Most studies still rely on strong human oversight and pedagogical revision.
Feedback on practice	4	C007, C008, C013, C019	Feedback-centered use cases are important but narrower than inquiry-scaffolding use cases.
Explicit epistemic understanding	2	C014, C019	Explicit epistemic work remains limited despite its relevance to science education.
Prompt transparency reported	8	C002, C003, C007, C008, C009, C012, C014, C019	Reporting quality is uneven; only part of the literature clearly describes prompt design.

Taken together, these patterns suggest that the current literature is stronger on initiating or structuring inquiry than on sustaining deeper forms of evidence-based disciplinary work. The empirical base shows more attention to planning, explanation, feedback, and communication than to argumentation, data interpretation, or explicit epistemic reflection. In practical terms, the current research emphasis is on making inquiry easier to start and manage, not yet on showing that GenAI can reliably deepen the quality of science-practice participation.

One plausible interpretation of this imbalance is that the currently dominant GenAI use cases map more easily onto front-end inquiry tasks than onto back-end disciplinary tasks. Planning, drafting explanations, or generating initial options are comparatively easier to scaffold with fluent language output, whereas sustained data interpretation, argumentation, and epistemic critique place greater demands on accuracy, multi-step reasoning, and resistance to plausible but weak justifications. Although the present corpus does not allow a causal claim about model architecture, the distribution of studies suggests that researchers may be gravitating toward practice dimensions that are easier to operationalize with current tools and lower-risk to implement in classrooms.

Reported Outcomes, Risks, and Research Gaps

The reported outcomes were mixed but informative. Across the corpus, cognitive or learning-related outcomes were reported in 9 studies, teacher- or design-oriented outcomes in 10 studies, assessment-related outcomes in 6 studies, epistemic outcomes in 5 studies, and affective outcomes in 4 studies. These categories overlapped, but their distribution still suggests that the literature is presently more comfortable reporting practical design and task-performance consequences than documenting deeper epistemic or identity-related change.

Several studies described productive uses of GenAI for inquiry-related support, including guided inquiry facilitation (Chang et al., 2023), improvement of density-concept learning through lesson-study integration (Cheng, 2025), support for causal modeling in environmental science contexts (Nguyen et al., 2025), and stronger support for open-ended experiment design in middle school physics (Min et al., 2025). Assessment-oriented studies also showed practical promise. Meyer et al. (2025), for example, reported that an open-source LLM achieved 0.84 classification accuracy and produced suitable feedback in 69% of cases on physics tasks, suggesting that school-controlled automated feedback may be feasible in at least some bounded domains. Chang et al. (2026) similarly suggested that dialogic AI-supported scoring and feedback can be useful in inquiry-instrument contexts when paired with human benchmarking.

At the same time, several studies identified important limitations. Stadler et al. (2024) found that LLM use reduced mental effort while also weakening the depth and quality of students' scientific justification. Haraldsrud and Odden (2025) similarly showed that productive use depended on students maintaining control over the modeling process rather than outsourcing reasoning to the AI. Teacher-facing lesson-planning studies revealed another recurring problem: GenAI could generate lesson plans that looked inquiry-oriented on the surface but drifted toward content transmission or inconsistent inquiry coherence when examined more closely (Kahraman & Kıyıcı, 2025; Großmann et al., 2025). These findings indicate that apparent fluency and convenience do not automatically translate into stronger science-practice engagement.

The risk pattern was also revealing. Implementation constraints were mentioned in all 14 studies, making them the most universal challenge in the corpus. Concerns linked to ethics, privacy, trust, or safety appeared in 8 studies, and issues related to epistemic authority or uncritical acceptance of AI output appeared in 7 studies. Accuracy concerns were raised in 5 studies, overreliance in 3 studies, and explicit hallucination or bias/fairness concerns in 2 studies each. This pattern suggests that the main barriers in this literature are not limited to classic "wrong answer" problems. More pervasive concerns involve how GenAI shapes authority, judgment, classroom responsibility, and the quality of participation in scientific work.

Across the included studies, productive outcomes were most credible when four conditions were present: a bounded task, explicit human oversight, reasonably transparent prompting, and criteria that allowed the AI output to be checked or revised. Where these conditions were absent, the literature more often reported drift, superficial inquiry alignment, or reduced quality of justification. Overall, the current empirical base suggests that GenAI is most established as a support layer for inquiry preparation, feedback, and design work. Evidence remains thinner for claims that GenAI already supports richer disciplinary engagement in data analysis, argumentation, and explicit epistemic reasoning. This imbalance marks an important gap for future science education research.

Discussion

The findings of this review suggest that the current GenAI literature in science education is developing along a narrower pathway than broad technology narratives often imply. Across the included studies, GenAI was most often used to support inquiry preparation, task structuring, lesson design, assessment feedback, or bounded forms of scientific reasoning. In other words, the literature currently positions GenAI more as a support layer around inquiry than as a robust facilitator of full disciplinary participation. This matters because science education is not defined only by efficient access to information or faster instructional planning. It is also defined by the quality of students' engagement with evidence, explanation, modeling, argumentation, and epistemic judgment. When the corpus is read through that lens, the empirical base appears promising but still uneven.

This interpretation helps clarify the contribution of the present review in relation to prior syntheses. Earlier reviews have already shown that AI and GenAI are becoming visible across science teaching and learning contexts (Almasri, 2024; Cheung et al., 2025), and JESEH has already published focused reviews on student learning outcomes and biology education (Duran & Dikmenli, 2025; Gunsaldi et al., 2025). However, those review lines do not fully answer a more discipline-sensitive question: whether GenAI is supporting the central practices through which science is learned and justified. The present review suggests that the answer is only partial. The literature is stronger on inquiry-adjacent support functions such as planning, feedback, and guided prompting than on richer practice participation such as argumentation, data interpretation, and explicit epistemic reflection. This pattern is also consistent with recent concerns that AI tools may support engagement or efficiency more readily than deeper disciplinary epistemic work (Cheung, 2026; Lee et al., 2026).

One useful way to interpret the corpus is to think of GenAI as infrastructural rather than transformational. In most included studies, the technology helped organize, annotate, draft, score, or suggest next steps. That is not trivial support; in some cases it clearly reduced workload, accelerated lesson design, or made feedback more available. Yet these functions are different from demonstrating that GenAI can deepen how learners weigh evidence, justify claims, challenge assumptions, or participate in scientific argumentation. The review therefore points to a field that is discovering where GenAI can be inserted into science-education workflows but is still far from proving that those insertions reliably strengthen the disciplinary core of science learning.

One of the clearest implications concerns teacher mediation. Even in studies that reported productive outcomes, GenAI rarely functioned well as an independent pedagogical actor. Inquiry chatbots required teacher monitoring when interactions broke down (Chang et al., 2023). Lesson-planning studies showed that AI-generated materials often needed substantial pedagogical revision before they could genuinely support inquiry (Kahraman & Kıyıcı, 2025; Großmann et al., 2025). Modeling and inquiry studies similarly indicated that productive use depended on whether learners retained control over the reasoning process rather than outsourcing key decisions to the model (Haraldsrud & Odden, 2025; Stadler et al., 2024). Taken together, these findings suggest that teacher judgment and instructional design remain central. The pedagogical question is therefore not whether GenAI can replace inquiry teaching, but under what conditions it can be orchestrated in ways that preserve scientific agency, scrutiny, and evidence-based reasoning.

This point has direct implications for science teacher education and professional development. If GenAI is to be used responsibly in science classrooms, teachers will need more than generic prompt-writing skills. They will also need practice-sensitive evaluative expertise: the ability to judge whether an AI-generated explanation supports conceptual sense-making, whether an inquiry plan genuinely preserves opportunities for investigation and revision, whether a chatbot response invites evidence-based reasoning rather than shortcut acceptance, and whether feedback aligns with disciplinary goals rather than only surface correctness. The teacher-facing studies in this corpus therefore imply that prompt literacy should be coupled with disciplinary and pedagogical review literacy. In pre-service teacher education, this could be translated into structured lesson-analysis tasks in which teacher candidates critique AI-generated science materials for inquiry coherence, epistemic quality, and alignment with science practices before adapting them for classroom use.

The findings also suggest that not all GenAI use cases are equally mature. Assessment and feedback tasks appeared relatively strong in the corpus, especially when the target task was bounded, criteria were explicit, and outputs could be checked against established expectations. In contrast, more open-ended uses such as inquiry-oriented lesson planning or broad information-seeking during socioscientific inquiry showed greater instability. This distinction has practical importance for science educators. Near-term adoption may be more defensible in constrained tasks such as feedback generation, assessment critique, or structured questioning support, whereas open-ended inquiry uses may require more cautious implementation, stronger prompts, clearer pedagogical framing, and closer human review. The issue is not simply technical accuracy. It is also whether the AI-supported activity preserves the epistemic aims of science education.

The review also points to a methodological maturity problem in the field. Most included studies were exploratory, qualitative, design-based, document-analytic, or short-term implementation studies. Only a small subset used comparative designs, larger performance datasets, or benchmark-style evaluation against human ratings. This is understandable in a fast-moving area, but it limits the kinds of claims that can presently be made. We know more about where GenAI is being tried and what immediate affordances or tensions appear than about how GenAI changes classroom inquiry over time, how stable those effects are across contexts, or whether gains in efficiency come at the cost of disciplinary depth.

Several directions for future research follow from this pattern. First, more work is needed on science practices that remain underrepresented in the current corpus, especially argumentation, data analysis, evidence interpretation, and explicit epistemic reflection. Second, the field would benefit from more classroom-based longitudinal studies that move beyond short pilots, document how prompts are designed, and show how teachers mediate AI use over time. Third, future studies should more clearly report prompt structures, human guidance, and revision procedures so that GenAI-supported activities can be interpreted and replicated more responsibly. A useful next step would be the adoption of a simple prompt-reporting framework that captures the model used, prompt purpose, prompt wording or structure, revision steps, human oversight, and any output-filtering procedures. Fourth, there is a need for stronger attention to equity, transparency, and trust, particularly in settings such as laboratory work, environmental education, and assessment where the consequences of inaccurate or opaque output may be pedagogically significant. Finally, comparative work should examine not only whether GenAI improves performance, but also whether it changes the quality of disciplinary participation in science.

The review also has important limitations. Most importantly, the search in this project was conducted through OpenAlex, Crossref, and targeted manual discovery rather than through a full subscription-database search in Scopus, Web of Science, and ERIC. This route was reproducible and sufficient for building a focused scoping corpus, but it may not have included all eligible studies indexed in subscription databases. The corpus was also relatively small and heterogeneous, spanning different educational levels, science domains, task types, and methodological designs. In addition, the review focused on English-language journal articles and did not include conference papers, dissertations, or other literature. Finally, as a scoping review, the purpose of the study was to map the evidence base rather than to conduct a formal methodological quality appraisal or effect-size synthesis. The findings should therefore be interpreted as a structured map of an emerging field rather than as a definitive estimate of what GenAI does in science education.

Conclusion

This scoping review mapped how generative AI has been used to support scientific inquiry and science practices in science education across 14 empirical studies published from late 2022 to early 2026. The review shows that GenAI is most commonly positioned as an inquiry scaffold, design aid, or assessment-and-feedback support rather than as a fully developed partner for deeper disciplinary reasoning. Investigation planning, explanation, and inquiry scaffolding were more visible in the literature than argumentation, data analysis, and explicit epistemic work.

Overall, the literature suggests that GenAI has genuine potential in science education, but that its most defensible uses currently lie in bounded and well-mediated tasks. The field has not yet produced strong evidence that GenAI consistently supports the full range of science practices at the depth science education requires. For this reason, future research should move beyond general claims of usefulness and examine more closely how GenAI shapes disciplinary participation, epistemic agency, and the quality of inquiry in science classrooms.

Scientific Ethics Declaration

* The author declares that the scientific ethical and legal responsibility of this article published in JESEH journal belongs to the author.

* This study reviewed published literature and did not directly involve human participants. Therefore, ethics committee approval was not required.

Generative AI Disclosure

* Generative AI tools were used in a limited editorial capacity during the preparation of this manuscript, primarily to support English-language clarity, grammar, sentence flow, and readability. The author remained fully responsible for the intellectual content, interpretation, organization, and final wording of the manuscript, and carefully reviewed and revised all AI-assisted suggestions before submission.

Conflict of Interest

* The author declares no conflict of interest.

Funding

* This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

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

Almasri, F. (2024). Exploring the impact of artificial intelligence in teaching and learning of science: A systematic review of empirical research. *Research in Science Education, 54*, 977-997. https://doi.org/10.1007/s11165-024-10176-3

Chang, J., Park, J., & Park, J. (2023). Using an artificial intelligence chatbot in scientific inquiry: Focusing on a guided-inquiry activity using Inquirybot. *Asia-Pacific Science Education, 9*(1), 44-74. https://doi.org/10.1163/23641177-bja10062

Chang, J., Park, J., & Sim, J. Y. (2026). Harnessing generative AI to facilitate epistemic understanding of students in scientific inquiry. *Journal of Science Education and Technology*. Advance online publication. https://doi.org/10.1007/s10956-026-10298-5

Cheng, E. C. K. (2025). Leveraging generative AI in science lesson study: Transforming density concept instruction through ChatGPT integration. *International Journal for Lesson and Learning Studies, 14*(3), 215-236. https://doi.org/10.1108/IJLLS-11-2024-0277

Cheung, K. K. C. (2026). Science identity and AI chatbots: A systematic review of empirical studies in science education. *Journal of Science Education and Technology*. Advance online publication. https://doi.org/10.1007/s10956-026-10295-8

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,* 1–29. https://doi.org/10.1080/03057267.2025.2578091

Doğru, M. S., & Faulconer, E. K. (2026). ChatGPT as a virtual laboratory teaching assistant in undergraduate biology. *Research in Science Education*, 56, 379-399. https://doi.org/10.1007/s11165-025-10271-z

Duran, T., & Dikmenli, M. (2025). Use of artificial intelligence in biology education: A systematic review of literature. *Journal of Education in Science, Environment and Health, 11*(4), 314-333. https://doi.org/10.55549/jeseh.864

Großmann, L., Koberstein-Schwarz, M., Krüger, D., & Krell, M. (2025). Lesson planning with ChatGPT for inquiry-based biology instruction: A(I) roll of the dice? *International Journal of Science Education,* 1–20. https://doi.org/10.1080/09500693.2025.2567509

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

Haraldsrud, A., & Odden, T. O. B. (2025). Bridging artificial intelligence and real intelligence: Self-scaffolding computational modeling with generative AI in chemistry. *Journal of Chemical Education, 102*(10), 4255-4266. https://doi.org/10.1021/acs.jchemed.5c00657

Kahraman, N., & Kıyıcı, G. (2025). Evaluating the efficacy of AI-generated inquiry-based lesson plans in science. *Sakarya University Journal of Education, 15*(1), 40-53. https://doi.org/10.19126/suje.1463067

Kim, W. J. (2026). Teachers' use of generative artificial intelligence for designing science lessons in support of environmental science agency. *Research in Science Education, 56*, 203-222. https://doi.org/10.1007/s11165-025-10262-0

Lee, G., Yun, M., Zhai, X., & Crippen, K. (2026). Artificial intelligence in science education research: Current states and challenges. *Journal of Science Education and Technology, 35*, 110-127. https://doi.org/10.1007/s10956-025-10239-8

Meyer, A., Bleckmann, T., & Friege, G. (2025). Automatic feedback on physics tasks using open-source generative artificial intelligence. *International Journal of Science Education*, 1–26. https://doi.org/10.1080/09500693.2025.2499220

Min, T., Lee, B., & Jho, H. (2025). Integrating generative artificial intelligence in the design of scientific inquiry for middle school students. *Education and Information Technologies, 30*(11), 15329-15360. https://doi.org/10.1007/s10639-025-13410-1

Nguyen, H., & Hayward, J. (2025). Applying generative artificial intelligence to critiquing science assessments. *Journal of Science Education and Technology, 34*, 199-214. https://doi.org/10.1007/s10956-024-10177-x

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

Peters, M. D. J., Godfrey, C., McInerney, P., Munn, Z., Tricco, A. C., & Khalil, H. (2020). Chapter 11: Scoping reviews. In E. Aromataris & Z. Munn (Eds.), *JBI Manual for Evidence Synthesis*. JBI. https://doi.org/10.46658/JBIMES-20-12

Peters, M. D. J., Marnie, C., Colquhoun, H., Garritty, C. M., Hempel, S., Horsley, T., Langlois, E. V., Lillie, E., O'Brien, K. K., Tunçalp, Ö., Wilson, M. G., Zarin, W., & Tricco, A. C. (2021). Scoping reviews: Reinforcing and advancing the methodology and application. *Systematic Reviews, 10*. https://doi.org/10.1186/s13643-021-01821-3

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior, 160*. https://doi.org/10.1016/j.chb.2024.108386

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., ... Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine, 169*(7), 467-473. https://doi.org/10.7326/M18-0850

Author(s) Information

M. Sait Gokalp
Kutahya Dumlupinar University
Evliya Çelebi Kampüsü. Tavşanlı Yolu 10. Km. Kütahya/Türkiye
Contact e-mail: *sgokalp@gmail.com*
ORCID iD: https://doi.org/0000-0002-3863-539X