Systematic Literature Review Protocol

1. Title of the study

Use of Generative Artificial Intelligence in Secondary Science Education and Teacher Training: A Systematic Literature Review

2. Authors of the study

Margarida M. Marques, Juliana Monteiro, Betina Lopes, Isabel Saúde, José Luís Araújo & J. Bernardino Lopes

3. Study description

This study is a mixed methods systematic literature review that investigates the use of Generative Artificial Intelligence (GAI) in Science Education and Teacher Training. It aims to review existing empirical literature (research papers and book chapters) to understand the theoretical foundations and practices related to GAI integration in education across key stage 7 to key stage 12 contexts, emphasizing science education and teacher training. The focus is on if and how educational stakeholders (students, teachers, future teachers and teacher educators) are using GAI for learning and teaching science.

This review includes quantitative, qualitative, and mixed method studies, grounded in the epistemological premise that integrating diverse methods enhances the depth and contextual relevance of findings. In light of the complex and emerging nature of GAI use in secondary science education and teacher training, a mixed-methods logic is adopted to identify patterns and outcomes, while also uncovering the mechanisms and contextual factors that shape implementation and impact, an approach supported by Jimenez et al. (2018). This option enables a more comprehensive understanding of how, why, and under what conditions GAI is being integrated into science teaching and learning practices.

4. Research question

What evidence does the literature provide on the use of Generative Artificial Intelligence (GAI) in teaching and learning practices within secondary science education (Years 7–12), involving students, in-service teachers, pre-service teachers, or teacher educators?

This question follows PICo (Participants, phenomenon of Interest and Context) framework (Hosseini et al. 2024).

Participants: individuals involved in secondary education (years 7–12), including students, in-service teachers (either in training or not), pre-service teachers, and teacher educators.

Phenomenon of interest: application of GAI in teaching and learning practices, encompassing both implementation and discourse about these practices, and their theoretical foundations.

Context: science subjects, such as biology, chemistry, geology, and physics, information technologies, and integrated STEM/STEAM initiatives that include science, either in formal or non-formal education contexts.

5. Study design

The mixed methods systematic literature review will follow the PRISMA declaration (Page et al., 2021) for transparent, complete, and accurate reporting. The review includes qualitative, quantitative, and mixed-methods empirical studies published from 2015 to April 2025.

6. Search strategy

Searches are conducted in Scopus, Web of Science core collection, with adapted strings.

Scopus search string:

TITLE-ABS-KEY ((AI OR "generative artificial intelligence" OR "generative AI" OR "Gen AI" OR GenAI OR chatbot OR "generative model" OR "large language model" OR LLM OR "natural language processing" OR NLP OR GPT OR copilot OR OpenAI OR Gemini OR Bard OR Llama OR Claude OR "DALL-E" OR Midjourney OR "Stable Diffusion" OR Imagen OR Gen OR "Leonardo.AI" OR "Leonardo AI" OR Veo OR "Pika labs" OR Sora OR Kaiber OR Lumen)
AND (educat* OR teach*)
AND ((educat* OR teach* OR learn* OR plan* OR instruct* OR curricul* OR didactic* OR resource OR material OR tutor OR mentor OR pedagogue OR student OR pupil OR

undergraduat* OR school* OR lesson OR class* OR lab* OR "formal education" OR "formal context") OR (tpack OR "Technological Pedagogical Content Knowledge" OR tam OR "Technology Acceptance Model" OR samr OR "Substitution, Augmentation, Modification and Redefiniton" OR theor*))
AND (science OR biology OR geology OR chemistry OR physics OR steam OR stem OR "information technologies" OR "information technology" OR ict OR "computer science")
AND ("School-year 7" OR "School-year 8" OR "School-year 9" OR "School-year 10" OR "School-year 11" OR "School-year 12" OR "seventh grade" OR "eighth grade" OR "ninth grade" OR "tenth grade" OR "eleventh grade" OR "twelfth grade" OR k12 OR "secondary school" OR "secondary education" OR "high school" OR "middle school" OR "teacher educat*” OR "pre-service teacher" OR "future teacher"))
AND PUBYEAR > 2014 AND PUBYEAR < 2026
AND (LIMIT-TO (LANGUAGE , "English") OR LIMIT-TO (LANGUAGE , "Spanish") OR LIMIT-TO (LANGUAGE , "German"))
AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "ch"))
Filters: ranges 2015-2015; document type: article and book chapter; language: English, Spanish, German, Portuguese

Web of science - Core Collection search string:

TS=((AI OR "generative artificial intelligence" OR "generative AI" OR "Gen AI" OR GenAI OR chatbot OR "generative model" OR "large language model" OR LLM OR "natural language processing" OR NLP OR GPT OR copilot OR OpenAI OR Gemini OR Bard OR Llama OR Claude OR "DALL-E" OR Midjourney OR "Stable Diffusion" OR Imagen OR Gen OR "Leonardo.AI" OR "Leonardo AI" OR Veo OR "Pika labs" OR Sora OR Kaiber OR Lumen)
AND (educat* OR teach*)
AND ((educat* OR teach* OR learn* OR plan* OR instruct* OR curricul* OR didactic* OR resource OR material OR tutor OR mentor OR pedagogue OR student OR pupil OR undergraduat* OR school* OR lesson OR class* OR lab* OR "formal education" OR "formal context") OR (tpack OR "Technological Pedagogical Content Knowledge" OR tam OR "Technology Acceptance Model" OR samr OR "Substitution, Augmentation, Modification and Redefiniton" OR theor*))
AND (science OR biology OR geology OR chemistry OR physics OR steam OR stem OR "information technologies" OR "information technology" OR ict OR "computer science")
AND ("School-year 7" OR "School-year 8" OR "School-year 9" OR "School-year 10" OR "School-year 11" OR "School-year 12" OR "seventh grade" OR "eighth grade" OR "ninth grade" OR "tenth grade" OR "eleventh grade" OR "twelfth grade" OR k12 OR "secondary school" OR "secondary education" OR "high school" OR "middle school" OR "teacher educator” OR "teacher education ” OR "pre-service teacher" OR "future teacher"))

AND PY=(2015-2025)
AND LA=("English" OR "Spanish" OR "German" OR "Portuguese")
AND DT=("Article" OR "Book Chapter")

7. Inclusion and exclusion criteria

The inclusion and exclusion criteria were explicitly defined according to the PICo structure of this study's research question.

Table 1. Inclusion and exclusion criteria for the selection of studies on the use of GAI in science education.

Type of criteria	Inclusion criteria	Exclusion criteria
a. Type of study	Empirical (qualitative, quantitative, or mixed methods)	Not empirical (e.g., exclude reviews, reflections)
b. Study participants	Students, in-service teachers, teachers in continuous training, pre-service teachers, or teacher educators involved in secondary education (Years 7–12)	Students below Year 7 or beyond Year 12, except when involved in teacher education programs; learners in other contexts
c. Phenomenon of interest / Study focus	Application of GAI in teaching and learning practices or the discourse about those practices, and their theoretical foundations	Not the use of GAI for teaching and learning (e.g., exclude learning analytics, non-generative AI chatbots like Siri)
d. Context	Science subjects (biology, geology, chemistry, physics), information technology, and STEM/STEAM initiatives that include science	Not explicitly situated within science education (e.g., exclude only mathematics, language, arts, programming, robotics)

8. Screening procedures, bias control and critical appraisal

After conducting the searches in the databases, the data is exported in ".RIS" files and subsequently imported to Rayyan.AI for screening management.

Duplicates are removed with the support of Rayyan.AI, in a semi-automated and user-mediated approach. Rayyan automatically compares bibliographic elements, such as titles, authors, etc., to identify records that are potentially redundant and generates a list of possible duplicates. Collaborative manual intervention from two researchers is conducted to confirm and resolve duplicates.

The screening procedures were initially drafted by two members of the research team and subsequently reviewed and refined through collective discussion, allowing all authors to propose adjustments until full consensus was achieved.

The studies will be screened through three phases, as described below. In phases 1 and 2, the screening process of each paper will be independently conducted by two researchers in a blind review to minimize the risk of bias. In these two phases, the papers are distributed among the reviewers, so that at least two reviewers are assigned to each article.

- Phase 1: Initial screening of title, abstract and keywords

Initially, the title, abstract and keywords of each paper will be screened to classify the publication as “include” or “exclude”, according to the agreed inclusion and exclusion criteria (see section “7. Inclusion and exclusion criteria"). Following this analysis, the researchers will meet to decide on classification conflicts. If needed, the inclusion and exclusion criteria are refined for increased clarity. In case of doubt, reviewers must decide to “include” the paper for full text screening in the second phase.

- Phase 2: Full text screening

Initially, each paper will be classified as “include” or “exclude”, according to the revised agreed inclusion and exclusion criteria and the full paper analysis. Following, the researchers will meet again to decide on classification conflicts and reach full consensus regarding the final analysis corpus.

- Phase 3: Critical appraisal

The methodological quality of the studies included in the analysis corpus will be appraised using the Mixed Methods Appraisal Tool (MMAT) developed by Hong and colleagues (2018). This tool offers a framework for assessing diverse empirical study designs, specifically focusing on their methodological quality. The study design categories considered in this tool are: a) Qualitative studies, b) Quantitative randomized controlled trials, c) Quantitative non-randomized studies, d) Quantitative descriptive studies, or e) Mixed methods studies.

To support the accurate classification of study designs, the MMAT user guide (Hong et al., 2018) provides descriptions of common research approaches within each category, along with a decision algorithm, which will serve as a reference during the appraisal process.

For studies classified as either qualitative or quantitative, only the corresponding single category will be applied. In the case of mixed methods studies, three sets of criteria are assessed: the qualitative component, one appropriate quantitative component, and the integrative mixed methods component. Within each component, each criterion will be rated using one of three response options: “Yes”, “No” or “Can’t tell”. Reviewers shall use the “Comments” column to succinctly justify each rating decision. In this study, reviewers will not consider the two initial screening questions, as the selection criteria for this review includes only empirical studies.

For the purposes of this review, the number of criteria marked as “Yes” will be used to generate a score for each study. Qualitative, Quantitative randomized controlled trials, Quantitative non-randomized, and Quantitative descriptive studies can achieve a maximum of 5 "Yes”, whereas Mixed methods studies may have a maximum of 15 "Yes”.

Two members of the research team will meet to collectively examine the MMAT criteria for each study design category, and to determine their appropriate application within the scope of this review. To foster a shared understanding, the reviewers will collectively apply the MMAT to five articles, one from each study design category. Interpretation discrepancies and uncertainties will be discussed until consensus is reached.

Subsequently, the remaining studies will be divided by the two reviewers to conduct independent appraisals. Finally, a new meeting will allow the reviewers to discuss doubts and challenging appraisals, as needed, to refine and agree on the final evaluations.

The results of the critical appraisal will be synthesized in a table listing the studies and the respective MMAT appraisal.

9. Data extraction and synthesis plan

Data extraction procedures were proposed by two authors of the review and analysed by all, who were able to propose adjustments, to avoid bias, until consensus was reached.

In an initial meeting, the reviewers will collectively extract data from a study, to reach a shared understanding of the data extraction form and its use within this review.

The remaining papers will be distributed among six reviewers so that at least two reviewers analyse each article. Each reviewer will assess the included studies independently, as described below.

Step 1. A preliminary (or exploratory) reading will be conducted for familiarization with the content of each article and to identify any internal inconsistencies or contradictions.

Step 2. A detailed and systematic reading will follow, during which data will be extracted using a structured acquisition form, as presented in Table 2. All log information must be explicit in the papers. Whenever inferences are made, they must be registered by adding the term “ - inferred” to the log. Whenever inference is not possible, the log must contain the term “Not specified”.

Table 2. List of data extraction thematic categories, data items, their description and justification within this literature review.

Thematic categories	Data items	Description	Justification
Publication and Contextual Metadata	Author(s)	All authors, in the form: Last name, initials	For citation
	Year of Publication	Year of study publication	For citation and to analyze trends over time
	Country/Region	Geographic location of the study	For cross-cultural/ contextual indicators
Study Design and Methodological Characteristics	Research design	Qualitative, quantitative, and mixed methods	For methodological synthesis
	Research approach	Approach explicit in the study (e.g., ethnography, survey, case study).	For methodological synthesis
	Data collection method(s) and validation	Interviews, surveys, experiments, document analysis, etc., and respective validations efforts.	For methodological synthesis
	Analysis method(s) and validation	Thematic analysis, statistical analysis, etc., and respective validations efforts.	For methodological synthesis
	Number of participants	Number	To assess study scope
	Participant type	Students, pre-service teachers, in-service teachers, teacher educators	To compare stakeholders' experiences and perspectives
	Educational level	K7–K12, tertiary/teacher education	To assess study scope
	Subject area (of the intervention or of the participants profile)	Biology, Physics, Chemistry, STEAM, ICT, etc.	To map GAI use across science domains
Variables of GAI Integration in Science Education and Teacher Training Note: if compared with another approach, make the differences explicit in each item of this dimension	Theoretical and conceptual framework(s) used	Explicit mention of TPACK, SAMR, TAM, constructivism, etc. (list all mentioned)	To map theoretical underpinnings of GAI integration in practices
	GAI tools	ChatGPT, Bard, Midjourney, DALL·E, etc. (and url, if not common)	To map GAI tools
	Type of GAI application in interventions	Content generation, feedback, curriculum development, etc.	To map GAI applications
	GAI-based teaching and learning practices in interventions	Brief description of the strategies, activities and/or other resources mobilized	To map GAI practices
	Enablers of GAI integration in interventions	Conditions identified that make easier to develop GAI-based interventions (technical, ethical, cognitive, or pedagogical)	For developing recommendations for policy and practice

Thematic categories	Data items	Description	Justification
	Barriers to GAI integration in interventions	Conditions identified that make more difficult to develop GAI-based interventions (technical, ethical, cognitive, or pedagogical)	For developing recommendations for policy and practice
Findings and implications	Participants' perceptions	Participants' views on GAI use in this context	To understand adoption factors and develop recommendations for policy and practice
	Impact of GAI-based teaching and learning	Empirically evaluated learning outcomes and other educational impact (both positive and negative impacts) in interventions	To assess the impact and develop recommendations for policy and practice
	Reflections and recommendations by study's authors	Insights and/or suggestions for future practice and/or research	To map insights and recommendations from the literature and gaps
Other unanticipated themes		Issues relevant in the study not anticipated prior to the analysis that may originate revision of the data items, indicating if they may be relevant under a previous thematic category.	To identify new relevant themes

As the aim is not to provide a meta-analysis of the results, the main study findings will be recorded qualitatively due to expected heterogeneity, and the results will be synthesized in a narrative review.

To ensure consistency, responses will be compared to identify agreement or any discrepancies. These will be discussed between reviewers and resolved until consensus is reached.

10.Anticipated limitations

Three limitations are anticipated.

First, the diverse and evolving definitions of GAI may complicate the screening process and, eventually, the synthesis of findings. This conceptual limitation may introduce bias and affect the clarity of our conclusions. This review study will adopt an explicit, clear and stable definition of GAI to mitigate this limitation.

Second, the review is susceptible to publication bias, meaning studies with positive results might be over-represented in the literature. This could lead to an overly optimistic view of GAI's applications and perceptions.

Finally, the rapid evolution of GAI technology presents a challenge to the timeliness of our findings. The fast pace of innovation means that some of the documented practices and tools may quickly become outdated, potentially limiting the long-term relevance of our conclusions.

11. Funding and conflicts of interest

This work is developed under the project GAI-SciTeach, which is funded by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the project 2023.13203.PEX (https://sciproj.ptcris.pt/176341PRJ).

The work of the first author is funded by FCT, under the Scientific Employment Stimulus - Individual Call (2022.02153.CEECIND; https://doi.org/10.54499/2022.02153.CEECIND/CP1720/CT0037).

The work of the second author is funded by FCT, under the research grant BIPD/UI57/12159/2025.

Conflicts of Interest: None declared

AI tools (ChatGPT-4 and Gemini) were used in the development of this document to assist with idea generation, grammar and style checks.

12. Dissemination plan

Findings will be published in a peer-reviewed education periodic, presented at international education conferences, and summarized for practitioner audiences.

13. References

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., & Nicolau, B. (2018). Mixed methods appraisal tool (MMAT), version 2018. *Registration of Copyright*, *1148552*(10), 1–7. http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/146002140/MMAT_2018_criteria-manual_2018-08-08c.pdf

Hosseini, M.-S., Jahanshahlou, F., Akbarzadeh, M. A., Zarei, M., & Vaez-Gharamaleki, Y. (2024). Formulating research questions for evidence-based studies. *Journal of Medicine, Surgery, and Public Health*, *2*, 100046. https://doi.org/10.1016/J.GLMEDI.2023.100046

Jimenez, E., Waddington, H., Goel, N., Prost, A., Pullin, A., White, H., Lahiri, S., & Narain, A. (2018). Mixing and matching: using qualitative methods to improve quantitative impact evaluations (IEs) and systematic reviews (SRs) of development outcomes. *Journal of Development Effectiveness*, *10*(4), 400–421. https://doi.org/10.1080/19439342.2018.1534875;REQUESTEDJOURNAL:JOURNAL:RJDE20;WGROUP:STRING:PUBLICATION

Page, M. J., Moher, D., 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., … Mckenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*, 1–36. https://doi.org/10.1136/BMJ.N160