Early Version – Uncorrected



AI-Generated Feedback for Nursing Documentation and Clinical Reasoning Training: A Scoping Review of Applications and Validation Approaches

Masarat Pariyanonth, Wannachai Wannasawade, Somkid Saelee

DOI: https://dx.doi.org/10.31584/jhsmr.20261495

Received: 5 February 2026

Revised: 17 April 2026

Accepted: 6 May 2026

Published online: 5 November 2026

Please cite this article as: Pariyanonth M, Wannasawade W, Saelee S. AI-generated feedback for nursing documentation and clinical reasoning training: a scoping review of applications and validation approaches. J Health Sci Med Res doi: http://dx.doi.org/10.31584/jhsmr.20261495

Early Version — Uncorrected: This accepted manuscript has not been typeset or proofread. Content and formatting may change prior to final publication and should not be regarded as the final Version of Record.

© 2026 JHSMR. Hosted by Prince of Songkla University.

This is an open access article under the CC BY-NC-ND license

(https://creativecommons.org/licenses/by-nc-nd/4.0/).

Original Article

AI-Generated Feedback for Nursing Documentation and Clinical Reasoning Training: A Scoping Review of Applications and Validation Approaches

Masarat Pariyanonth, RN, M.Sc., Wannachai Wannasawade, Ph.D., Somkid Saelee, Ph.D.

Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand.

Corresponding Author:

Masarat Pariyanonth, RN, MSc.

Department of Computer Education, Faculty of Technical Education King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand.

Email: masarat.p@email.kmutnb.ac.th

Early Version - Uncorrected

Abstract

Objective: This scoping review aimed to map the application characteristics of AI-generated feedback in nursing documentation and clinical reasoning training and to synthesize the validation approaches used to establish trustworthiness.

Material and Methods: Guided by the Arksey and O'Malley framework, a structured search was conducted across five databases for peer-reviewed studies published between 2021 and 2025. Fourteen studies met the eligibility criteria. Data were synthesized descriptively to characterize AI application features and validation practices.

Results: The included studies utilized diverse AI systems, including LLM-based platforms, chatbot tutors, retrieval-augmented generation (RAG) systems, automated scoring engines, and mixed-reality applications, delivering both real-time and post-task feedback across varied nursing learning contexts. Validation approaches were highly heterogeneous, revealing substantial gaps in the evidence of trustworthiness. While many studies demonstrated strong evidence in expert review, guideline/rubric alignment, and reliability testing, evidence for robustness across cases and safety safeguards remained limited.

Conclusion: The findings highlight a critical validation gap, in which technological adoption is advancing faster than the evidence supporting technical reliability, educational validity, and clinical safety trustworthiness. Future research should prioritize standardized validation frameworks that explicitly address safety-critical metrics to ensure the responsible implementation of AI-generated feedback in nursing education.

Keywords: AI-generated feedback, artificial intelligence, clinical reasoning, nursing documentation, nursing education

Early Version - Uncorrected

Introduction

Feedback is widely recognized as a pivotal mechanism through which learners refine performance, calibrate professional judgment, and internalize clinical standards, particularly in complex and iterative tasks. In nursing education, nursing documentation and clinical reasoning represent two such critical domains. Documentation serves as a safety-critical communication record essential for continuity of care and professional accountability, while clinical reasoning facilitates sound decision-making in dynamic clinical environments. Recent literature emphasizes that enhancing documentation quality is intrinsically linked to patient safety and care outcomes, especially within electronic health record systems[1]. Concurrently, nursing programs face growing structural challenges that limit their capacity to provide high-quality feedback at scale. These obstacles include large student enrollments, insufficient faculty staffing, and inconsistent clinical placement opportunities. These constraints often result in delayed, inconsistent, or non-specific feedback, conditions that potentially diminish learning gains in skills requiring deliberate practice and structured reflection. Consequently, the integration of Artificial Intelligence (AI) in nursing education has accelerated, encompassing applications such as simulation support, intelligent tutoring systems (ITS), and automated assessment and feedback tools[2,3].

Recent studies indicate that AI applications are being utilized across diverse pedagogical frameworks and settings, with the potential to enhance student engagement, instructional efficiency, and competency development. However, these advancements necessitate a critical examination of transparency, ethical implications, and governance frameworks[4,5]. Moreover, the advent of Generative AI (GenAI) and Large Language Models (LLMs) has shaped the way students compose and rationalize clinical records[6,7]. With the progression of technological capabilities, AI-facilitated feedback mechanisms have grown progressively complex, extending from immediate algorithm-driven cues to descriptive recommendations congruent with the clinical evaluation criteria[4,6].

The educational value of AI-generated feedback depends critically on its reliability and accuracy. Clinical documentation and reasoning in nursing require sophisticated judgment and contextual understanding rather than simple right-or-wrong answers. When AI feedback contains inaccuracies or fails to reflect established professional standards, it risks misleading students and

Early Version - Uncorrected

encouraging unwarranted confidence in their abilities. Research examining GenAI performance in nursing diagnosis has shown inconsistent accuracy levels, underscoring the need for thorough validation[8]. While AI-enhanced assessments show promise, questions persist regarding the quality of feedback and how students interpret and apply this guidance[9].

A major ongoing issue is the inconsistent approaches to validating AI feedback systems. Much of the existing literature emphasizes usability and user perceptions of usefulness[10]. However, few studies present comprehensive validation frameworks that address both technical and educational dimensions. Technical considerations include accuracy and reliability, while educational aspects encompass alignment with pedagogical goals and the practical applicability of feedback. Current reviews examining GenAI applications in nursing education reveal diverse implementations but also highlight problematic variations in evaluation methods and unresolved ethical considerations[6,11].

Given the emerging nature of this field, a scoping review is warranted to systematically map the existing evidence on the implementation of AI-generated feedback in nursing documentation and clinical reasoning training, clarify its applications, and synthesize how validation processes have been operationalized. Contemporary guidance for scoping reviews highlights the importance of robust protocols and transparent reporting, particularly in methodologically diverse research areas[12–14].

Accordingly, this review aimed to characterize the application features of AI-generated feedback, including AI techniques, feedback targets, timing, formats, and educational contexts, and to examine the validation approaches used to establish trustworthiness, such as technical accuracy, consistency, educational validity, and human-in-the-loop oversight. By synthesizing current practices and identifying gaps in validation, this review sought to inform future research and support the responsible integration of AI-generated feedback within nursing education.

Materials and Methods

1. Study design and reporting approach

This study was conducted as a scoping review guided by the methodological framework proposed by Arksey and O'Malley[15] to map the scope, characteristics, and validation approaches of AI-generated feedback used in nursing documentation and clinical reasoning training. A

Early Version - Uncorrected

structured search of peer-reviewed literature published within the past five years (January 2021 to December 2025) was performed using PubMed/MEDLINE as the primary bibliographic database and supplemented by searches in publisher platforms (ScienceDirect and Wiley Online Library) and academic search engines (Google Scholar and Semantic Scholar); records retrieved from Google Scholar and Semantic Scholar were screened to ensure inclusion of peer-reviewed journal articles only. All retrieved citations were imported into Rayyan (Rayyan Systems Inc.), which was used to facilitate the screening process, manage study selection, and identify duplicates. Additionally, the reference lists of the included studies were manually searched to identify further relevant publications.

Search terms included combinations of "artificial intelligence", "AI-generated feedback", "nursing documentation", and "clinical reasoning", with database-specific adaptations.

2. Review objectives and key concepts

The review had two objectives: (1) to map application characteristics of AI-generated feedback (AI techniques, feedback targets, timing, format, and educational context), and (2) to synthesize the validation approaches used to establish the trustworthiness of AI-generated feedback, including technical evidence, educational validity, and the presence of human-in-the-loop oversight.

Definition of AI-generated feedback:

AI-generated feedback was defined as AI-produced information intended to help learners improve their performance in nursing documentation and/or clinical reasoning (e.g., rubric-aligned comments, suggestions, prompts, annotations, or scores), delivered either in real-time or post-task. Studies were considered in-scope only when feedback was directed at learner performance outputs or decisions rather than limited to information provision, content delivery, or general conversation without performance-focused feedback.

3. Eligibility criteria

Eligibility was defined using the PCC framework, focusing on nursing learners (Population), AI-generated performance feedback for documentation and clinical reasoning (Concept), and nursing education or training settings (Context). Studies were included if they reported AI-generated feedback directed at learner performance in nursing documentation and/or

Early Version - Uncorrected

clinical reasoning within nursing education contexts and excluded if AI was used solely for information provision or general conversation without performance-focused feedback.

4. Study selection

Records were screened in two stages by two independent reviewers: title/abstract screening followed by full-text assessment against the eligibility criteria. Eligibility decisions for ambiguous cases were resolved through discussion and consensus, and when necessary, a third reviewer was consulted. Reasons for exclusion at the full-text stage were recorded to ensure transparency.

5. Data charting and variables

Data were charted independently by two reviewers using a structured extraction form and managed in Microsoft Excel. Extracted data included study characteristics, application features, and validation evidence. Validation was synthesized using a matrix approach that coded evidence across predefined domains on a 0 – 2 scale (0 = not reported, 1 = partial/unclear, 2 = clearly reported). The validation domains were derived from prior literature and reviewed by experts in nursing education and AI to ensure content validity.

In this review, educational validity was operationalized as evidence that AI-generated feedback supports intended educational uses, including alignment with learning outcomes or standards. Trustworthiness was treated as a multi-component construct encompassing technical accuracy or consistency, educational validity, and safeguards or oversight.

6. Synthesis of results

Findings were synthesized descriptively. Study characteristics were summarized using counts and tables. Application patterns were mapped by AI techniques, feedback timing and format, and targeted competency areas. Validation approaches were summarized by domain frequency and narrative comparison of how trustworthiness, including technical and educational validation evidence, was established.

No meta-analysis was planned, as the purpose of this scoping review was to map the characteristics and validation approaches of existing evidence rather than to generate pooled effect estimates. Accordingly, a formal quality appraisal of included studies was not undertaken, in line with scoping review methodology.

Early Version - Uncorrected

7. Ethics statement

Ethical approval was not required because this study synthesized data from published literature and did not involve human participants or identifiable personal data.

Results

1. Study selection and overview

The study selection process is illustrated in Figure 1. After removing duplicates, titles and abstracts were screened for relevance. Full-text assessments were then conducted based on the predefined eligibility criteria, resulting in the inclusion of 14 studies in the final review.

Figure 1 Study selection flow diagram

2. Characteristics of the included studies

As summarized in Table 1, the included studies were conducted across diverse geographical regions and educational settings, primarily involving undergraduate nursing students and, in some cases, postgraduate trainees. Study designs varied and included pilot evaluations, quasi-experimental studies, and descriptive or mixed-methods approaches. The AI technologies employed ranged from rule-based systems to natural language processing–based tools and large language model–driven applications. The educational focus of the studies encompassed nursing documentation tasks, care planning, and clinical reasoning activities embedded within simulations, case-based learning, or digital learning environments.

Table 1 Characteristics of the included studies on AI-generated feedback for nursing documentation and clinical reasoning training

3. Applications of AI-generated feedback

The characteristics of AI-generated feedback applications are detailed in Table 2. Across the included studies, AI systems ranged from LLM-based platforms (e.g., ChatGPT, GPT-4, RAG systems) to chatbot tutors, automated essay-scoring engines, and mixed-reality simulation platforms. AI-generated feedback was applied across diverse instructional contexts, including simulation, online learning, in-class activities, and hybrid formats, with simulation-based

Early Version - Uncorrected

scenarios being the most common. Clinical scenarios covered a broad range of nursing domains, such as emergency response, psychiatric nursing, maternal-child care, labor nursing, and therapeutic communication training.

Feedback outputs varied substantially across systems, including automated scoring, rubric-aligned evaluations, structured debriefing summaries, individualized prompts, corrective hints, and AI-generated recommendations for nursing documentation. Feedback timing also differed across studies: some systems provided immediate real-time responses during learner interaction, whereas others delivered post-task feedback after task completion; several studies incorporated both modes. Most applications employed adaptive feedback mechanisms, often involving personalized or context-sensitive responses tailored to learner input, although the degree of adaptivity was not always explicitly specified. Overall, these findings demonstrate considerable diversity in how AI-generated feedback is operationalized to support nursing documentation and clinical reasoning training.

4. Validation approaches and trustworthiness

Validation approaches used to assess AI-generated feedback are summarized in Table 3 and were highly heterogeneous. Across studies, reported evaluations commonly emphasized feasibility-oriented outcomes (e.g., user experience and acceptance), whereas fewer studies provided rigorous evidence directly related to feedback trustworthiness, such as benchmarking against explicit standards, agreement with expert ratings, or tests of reliability and consistency. Evidence of educational validity was inconsistently reported, with variable descriptions of alignment with learning outcomes and the extent to which feedback supported meaningful learner improvement. Descriptions of oversight or safeguard mechanisms were also uncommon. Collectively, these findings indicate persistent gaps in the systematic and transparent validation of AI-generated feedback for nursing documentation and clinical reasoning training.

It should be noted that core validation indicates the presence of at least one form of explicit validation evidence and does not imply comprehensive or exhaustive validation across all domains. Accordingly, studies classified as having core validation may still lack evidence in other critical areas, such as safety or robustness. Thus, core validation should be interpreted as a minimal threshold rather than an indicator of overall validation quality.

Early Version - Uncorrected

The number of strong validation domains ranged from one to five across studies. One study had one strong domain; no study had two; three had three; six had four; and four had five strong domains (Figure 2). Notably, no study provided strong evidence across all six validation domains, indicating uneven but generally moderate validation strength across the included studies.

Table 2 Application characteristics of AI-generated feedback in nursing documentation and clinical reasoning training

Table 3 Validation approaches and trustworthiness evidence for AI-generated feedback systems (0–2 scoring per domain)

Figure 2 Distribution of the Included Studies by Number of Strong Validation Domains (maximum possible = 6)

Strong validation domains refer to validation domains in which explicit evidence was reported (score = 2). No study demonstrated strong evidence across all six validation domains.

Discussion

This scoping review systematically mapped the applications and validation approaches of AI-generated feedback systems in nursing documentation and clinical reasoning training. The findings reveal a rapidly evolving landscape characterized by the adoption of sophisticated AI models, such as Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) systems, alongside a notable and concerning heterogeneity in validation practices. The synthesis of evidence highlights a notable gap between the technological advancement of these systems and the rigor applied to establishing their trustworthiness, particularly concerning patient safety and educational validity.

The Validation Gap: Trustworthiness Versus Innovation

A central finding of this review is the inconsistent and, in many cases, limited validation of AI-generated feedback systems, particularly in domains critical to patient care. While the included studies suggest an increasing use of advanced and adaptive AI techniques (Table 2), the corresponding evidence supporting core trustworthiness is often less robust. Specifically, analysis of validation approaches (Table 3) indicates that domains related to safety/clinical risk checks

Early Version - Uncorrected

were most often reported with limited detail (predominantly score = 1), while reliability and consistency across cases, although more frequently addressed, still showed uneven rigor across studies. Notably, this imbalance was most evident in the contrast between validation domains: while guideline- or rubric-based benchmarking was strongly represented in nearly all studies, explicit safety or clinical risk validation was reported in only a small minority. This pattern suggests that current validation practices prioritize alignment with educational standards over safeguards against clinically unsafe or misleading AI-generated feedback. The distribution of strong validation domains (Figure 2) demonstrates moderate heterogeneity in validation practices, with most studies clustering at four strong domains, while fewer studies achieved either minimal or highly comprehensive validation coverage. Importantly, no study demonstrated strong evidence across all six validation domains, underscoring the absence of fully comprehensive trustworthiness frameworks in current AI-generated feedback systems. This pattern suggests that although many AI-generated feedback systems have incorporated multiple forms of validation, important gaps remain in achieving a balanced, fully transparent assessment of trustworthiness across all critical domains.

This gap is particularly important given that nursing documentation and clinical reasoning are safety-critical competencies[1]. Feedback inaccuracies, or hallucinations, from LLM-based systems could potentially mislead students, fostering unwarranted confidence or, worse, embedding incorrect clinical standards[28]. While strong validation evidence was common in expert review, guideline/rubric alignment, and reliability testing (Table 3), these results typically confirm the system's design or initial alignment rather than its performance consistency across varied clinical scenarios. Therefore, the academic community must urgently shift its focus from merely demonstrating the feasibility and usability of AI feedback. It should instead establish comprehensive, transparent, and reproducible validation frameworks that prioritize patient safety and educational integrity[12,29].

The Adaptive Feedback Paradigm and Its Educational Implications

The widespread adoption of adaptive feedback mechanisms, often powered by LLMs and other intelligent AI systems (Table 2), represents a notable pedagogical shift from traditional rule-based systems. Adaptive feedback, characterized by descriptive and context-sensitive guidance

Early Version - Uncorrected

delivered in real-time, post-task, or combined formats, effectively mirrors the interaction style of human experts. This sophistication may substantially reduce students' cognitive load and support deliberate practice in complex skills such as clinical reasoning[30]. However, this shift introduces new educational considerations. The ease and immediacy of adaptive feedback may inadvertently lead to over-reliance or overconfidence in the AI's judgment, potentially diminishing the student's intrinsic motivation for deep, reflective learning[31]. Furthermore, the black-box nature of many LLMs challenges the principle of feedback transparency, making it difficult for students and educators to understand why a specific piece of feedback was generated. Future research must explore the optimal balance between AI-driven efficiency and the preservation of critical thinking and self-correction skills in nursing students.

Redefining the Educator's Role: Human-in-the-Loop Oversight

The findings suggest that AI-generated feedback systems are not replacements for nurse educators but rather tools that necessitate a redefinition of the educator's role. In the era of GenAI, the educator transitioned from being the primary source of feedback to becoming the "Validator" and "Facilitator" within a robust "human-in-the-loop" oversight model.

The high scores for the expert reviews in Table 3 confirm that human oversight is currently concentrated at the system development and content alignment stages. However, the low scores in the Safety/Clinical risk check suggest a need for continuous, real-time human monitoring and intervention during the system's operational use. Educators must be trained to identify and mitigate AI-generated inaccuracies, interpret the AI's output for students, and effectively integrate AI feedback within the broader clinical context. This transition requires dedicated professional development for faculty to ensure they possess the necessary digital literacy and critical appraisal skills to manage AI-enhanced learning environments effectively[4].

Heterogeneity and the Challenge of Standardization

The review highlights substantial heterogeneity across the included studies, encompassing diverse AI techniques (ranging from simple chatbots to RAG and Mixed Reality systems), instructional contexts (e.g., simulation, online learning), and geographical regions (Table 1). While this diversity reflects the global innovation in the field, it simultaneously poses a significant challenge to the standardization of quality and efficacy. The lack of a common metric

Early Version - Uncorrected

for educational validity and technical trustworthiness makes it difficult to compare the effectiveness of different AI systems or to generalize findings across various nursing programs[32]. Future progress in this domain necessitates standardized reporting protocols and validation benchmarks to ensure that AI feedback integration proceeds responsibly and is grounded in robust, reproducible empirical data.

When comparing outputs across AI system types represented in this review, meaningful differences emerge in feedback characteristics and validation rigor (Tables 2 and 3). LLM-based systems (e.g., GPT-4 and RAG-based platforms) enabled more flexible and context-responsive feedback generation, but also demonstrated wider variability in validation strength, with some studies achieving five strong validation domains (S09, S13, S14), while others reported more limited validation coverage (e.g., S10) (Table 3). In contrast, automated essay-scoring systems such as IntelliMetric® (S22) and structured chatbot platforms (S19, S21) tended to report more consistent reliability and rubric-alignment evidence, which may reflect their more structured and rule-governed scoring designs. Mixed-reality systems (S09) stood out by integrating situational awareness testing (SAGAT) alongside AI feedback, representing one of the most comprehensive multimodal validation approaches observed. These patterns reinforce the need for more rigorous validation as AI systems become increasingly generative and open-ended.

This review has several limitations. First, only selected databases were included (PubMed/MEDLINE, ScienceDirect, Wiley Online Library, Google Scholar, and Semantic Scholar), which may have resulted in the omission of relevant studies indexed in other databases, such as CINAHL and EMBASE. Second, the heterogeneity of the included studies limited direct comparison across findings. Third, although the validation approaches were systematically mapped, a formal quality appraisal was not conducted. This is consistent with the scoping review methodology, as the Joanna Briggs Institute (JBI) framework does not require critical appraisal as a mandatory component unless the review intends to make evaluative judgments about evidence quality. Future systematic reviews in this area should incorporate formal quality appraisal tools to enable stronger evidence synthesis.

Early Version - Uncorrected

Conclusion

The integration of AI-generated feedback into nursing education holds immense promise for addressing the challenges of scale and consistency in feedback delivery. However, this review demonstrates that the current evidence base is fragmented due to a lack of critical validation. To facilitate the responsible implementation of AI in this safety-critical domain, future research must prioritize the development and application of standardized, transparent validation frameworks that explicitly address technical accuracy, educational validity, and, most importantly, clinical safety. The success of AI in nursing education hinges not only on technological innovation but also on the commitment of researchers and educators to conduct rigorous, human-centered oversight.

References

1. Saraswasta IWG, Hariyati RrTS. A systematic review of the implementation of electronic nursing documentation toward patient safety. Enferm Clin 2021;31:S205–9. doi:10.1016/j.enfcli.2020.12.023.
2. Jung S. Challenges for future directions for artificial intelligence integrated nursing simulation education. Korean J Women Health Nurs 2023;29(3):239–42. doi: 10.4069/kjwhn.2023.09.06.1.
3. Mitha Z, Schwartz J, Hobensack M, Cato K, Woo K, Smaldone A, et al. Natural Language Processing of Nursing Notes: An Integrative Review. Comput Inform Nurs 2023;41(6):377–84. doi: 10.1097/CIN.0000000000000967.
4. Labrague LJ, AL Sabei S, AL Yahyaei A. Artificial intelligence in nursing education: A review of AI-based teaching pedagogies. Teach Learn Nurs 2025;20(3):210–21. doi: 10.1016/j.teln.2025.01.019.
5. El Arab RA, Al Moosa OA, Abuadas FH, Somerville J. The Role of AI in Nursing Education and Practice: Umbrella Review. J Med Internet Res 2025;27:e69881. doi: 10.2196/69881.
6. Park GE, Kim H, Go UR. Generative artificial intelligence in nursing: A scoping review. Collegian 2024;31(6):428–36. doi: 10.1016/j.colegn.2024.10.004.
7. Harrington J, Booth RG, Jackson KT. Large Language Models in Nursing Education: Concept Analysis. JMIR Nurs 2025;8:e77948. doi: 10.2196/77948.

Early Version - Uncorrected

8. Ju H, Park M, Jeong H, Lee Y, Kim H, Seong M, et al. Generative AI-Based Nursing Diagnosis and Documentation Recommendation Using Virtual Patient Electronic Nursing Record Data. Healthc Inform Res 2025;31(2):156–65. doi: 10.4258/hir.2025.31.2.156.
9. Wang XJ, Song LJ, Jiao XP, Chen SQ. Integration of Artificial Intelligence in Nursing Clinical Education: Enhancing the Mini-CEX Model. J Multidiscip Healthc 2025; 18:7327–37. doi: 10.2147/JMDH.S550145.
10. Han J won, Park J, Lee H. Development and effects of a chatbot education program for self-directed learning in nursing students. BMC Med Educ 2025;25(1):825. doi: 10.1186/s12909-025-07316-2
11. Tischendorf T, Hinsche L, Hasseler M, Schaal T. GenAI in nursing and clinical practice: a rapid review of applications and challenges. J Public Health (Berl.) 2025. doi: 10.1007/s10389-025-02523-z
12. Peters MDJ, Godfrey C, McInerney P, Khalil H, Larsen P, Marnie C, et al. Best practice guidance and reporting items for the development of scoping review protocols. JBI Evid Synth 2022;20(4):953–68. doi: 10.11124/JBIES-21-00242
13. Pollock D, Evans C, Menghao Jia R, Alexander L, Pieper D, Brandão de Moraes É, et al. "How-to": scoping review? J Clin Epidemiol 2024;176. doi: 10.1016/j.jclinepi.2024.111572
14. Hadie SNH. ABC of a Scoping Review: A Simplified JBI Scoping Review Guideline. Educ Med J 2024;16(2):185–97. doi: 10.21315/eimj2024.16.2.14
15. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19–32. doi: 10.1080/1364557032000119616.
16. Xuto P, Prasitwattanaseree P, Chaiboonruang T, Chaiwuth S, Khwanngern P, Nuntakwang C, et al. Development and Evaluation of an AI-Assisted Answer Assessment (4A) for Cognitive Assessments in Nursing Education. Nurs Rep 2025;15(3):80. doi: 10.3390/nursrep15030080
17. Olla P, Wodwaski N, Long T. Beyond the Bot: A Dual-Phase Framework for Evaluating AI Chatbot Simulations in Nursing Education. Nurs Rep 2025;15(8). doi: 10.3390/nursrep15080280.
18. Ching YC, Ho YC. AI Virtual Human–Augmented Game-Based Teaching to Enhance Emotional Intelligence in Nursing Students: Protocol for a Single-Group Pretest-Posttest Action Research Study. JMIR Res Protoc 2025;14:e80290. doi: 10.2196/80290.

Early Version - Uncorrected

19. Sepanloo K, Shevelev D, Islam MT, Son YJ, Aras S, Hinton JE. Improving nursing education through an AI-enhanced mixed reality training platform: development and pilot evaluation. Educ Technol Res Dev 2025;73(3):1835–63. doi: 10.1007/s11423-025-10473-2.
20. Arkan B, Dallı ÖE, Varol B. The impact of ChatGPT training in the nursing process on nursing students' problem-solving skills, attitudes towards artificial intelligence, competency, and satisfaction levels: Single-blind randomized controlled study. Nurse Educ Today 2025;152:106765. doi: 10.1016/j.nedt.2025.106765.
21. Jang A, Oh M, Song MO. Question-centered prompt engineering for nursing simulation debriefing: Model development and validation. Nurse Educ Pract 2025;88. doi: 10.1016/j.nepr.2025.104541.
22. Shen M, Shen Y, Shi Z. Exploring nursing students' acceptance of RAG-enhanced GenAI through the AIDUA model: A qualitative study. Nurse Educ Pract 2025;89. doi: 10.1016/j.nepr.2025.104612.
23. Kroll N, Thai L, Seo JH, Godbole MS, Weston C, Wells-Beede E. Generative AI in simulation-based SBIRT training: Enhancing content validity and educational impact. Clin Simul Nurs 2025;108. doi: 10.1016/j.ecns.2025.101811.
24. Han JW, Park J, Lee H. Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: a quasi-experimental study. BMC Med Educ 2022;22(1):830. doi: 10.1186/s12909-022-03898-3.
25. Fung TCJ, Chan SL, Lam CFM, Lam CY, Cheng CCW, Lai MH, et al. Effects of generative artificial intelligence (GenAI) patient simulation on perceived clinical competency among global nursing undergraduates: a cross-over randomised controlled trial. BMC Nurs 2025;17;24(1):934. doi: 10.1186/s12912-025-03492-0.
26. Park SA, Kim HY. Development and effects of a scenario-based labor nursing simulation education program using an artificial intelligence tutor: a quasi-experimental study. J Womens Health Nurs 2025;31(2):143–54. doi: 10.4069/whn.2025.06.18.
27. Wolf RR, Wolf A. Using AI to Evaluate a Competency-Based Online Writing Course in Nursing. Online Learn 2023;27(3):1-29. doi:10.24059/olj.v27i3.3974.

Early Version - Uncorrected

28. Lin Z, Guan S, Zhang W, Zhang H, Li Y, Zhang H. Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. Artif Intell Rev 2024;57(9):243. doi: 10.1007/s10462-024-10896-y
29. Kwong JCC, Khondker A, Lajkosz K, McDermott MBA, Frigola XB, McCradden MD, et al. APPRAISE-AI Tool for Quantitative Evaluation of AI Studies for Clinical Decision Support. JAMA Netw Open 2023;6(9):e2335377. doi: 10.1001/jamanetworkopen.2023.35377.
30. Akavova A, Temirkhanova Z, Lorsanova Z. Adaptive learning and artificial intelligence in the educational space. E3S Web of Conf 2023;451:06011. doi: 10.1051/e3sconf/202345106011
31. Tam W, Huynh T, Tang A, Luong S, Khatri Y, Zhou W. Nursing education in the age of artificial intelligence powered Chatbots (AI-Chatbots): Are we ready yet? Nurse Educ Today 2023;129:105917. doi: 10.1016/j.nedt.2023.105917.
32. Seibert K, Domhoff D, Bruch D, Schulte-Althoff M, Fürstenau D, Biessmann F, et al. Application Scenarios for Artificial Intelligence in Nursing Care: Rapid Review. J Med Internet Res 2021;23(11):e26522. doi: 10.2196/26522.

Early Version - Uncorrected

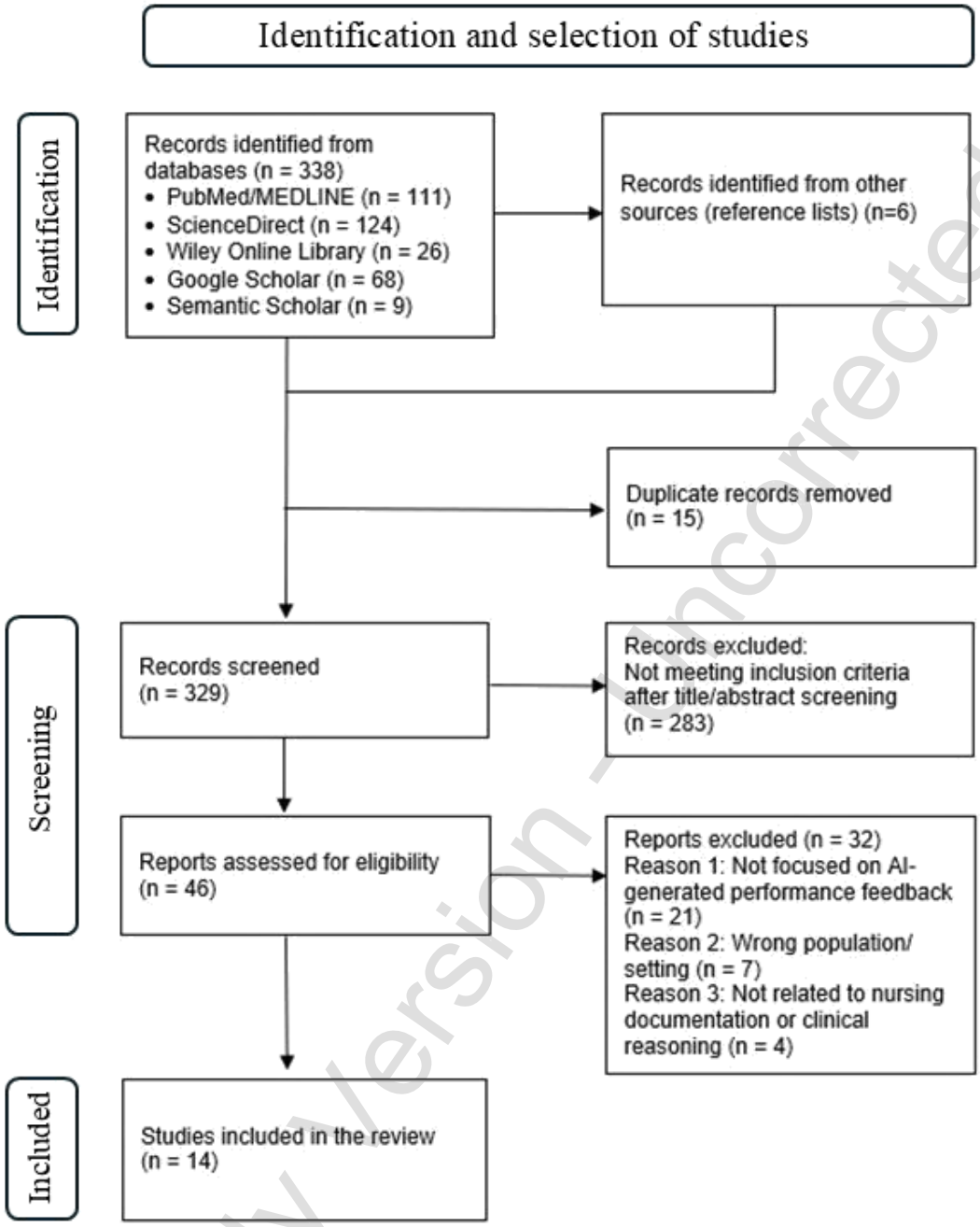


Figure 1 Study selection flow diagram

Early Version - Uncorrected

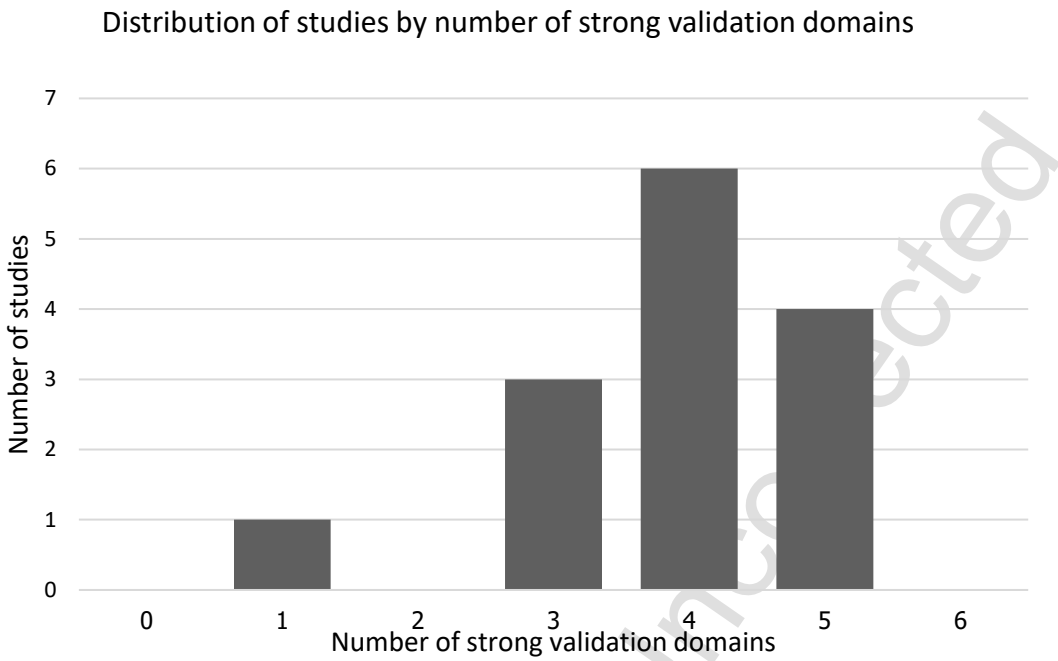


Figure 2 Distribution of the Included Studies by Number of Strong Validation Domains (maximum possible = 6)

Strong validation domains are those in which explicit evidence was reported (score = 2). No study demonstrated strong evidence across all six validation domains.

Early Version - Uncorrected

Table 1 Characteristics of the included studies on AI-generated feedback for nursing documentation and clinical reasoning training

Study ID	Authors (Year)	Country/Region	Study design	Setting	Participants (n)	AI tool/Model
S05	Ju et al. (2025)[8]	South Korea	Comparative pilot study	Simulation	Registered nurses (n=40)	ChatGPT 4.0
S06	Xuto et al. (2025)[16]	Thailand	Validation study utilizing cross-sectional data	Online	Undergraduate nursing students (n=308)	ChatGPT 4o/Speech-to-Text (STT)/
S07	Olla et al. (2025)[17]	United States	Simulation-based study	Simulation	Advanced practice nurses (n=3)	ChatGPT 4.0
S08	Ching&Ho (2025)[18]	Taiwan	Action research	In-class and Simulation	Postbaccalaureate nursing students (n=37)	LLM-based simulation system (ChatGPT)
S09	Sepanloo et al. (2025)[19]	United States	Pilot study	Simulation	Nursing students and nursing faculty (n=10)	Natural Language Processing (NLP) and the BERT model
S10	Arkan et al. (2025)[20]	Turkey	Randomized controlled trial	Online and In-class	Undergraduate nursing students (n=96)	ChatGPT 3.5

Early Version - Uncorrected

Study ID	Authors (Year)	Country/Region	Study design	Setting	Participants (n)	AI tool/Model
S13	Jang et al. (2025)[21]	South Korea	Mixed-methods study	Simulation	Undergraduate nursing students (n=28)	ChatGPT 4.0 (Prompt Engineering)
S14	Shen et al. (2025)[22]	China	Qualitative exploratory design	Online	Undergraduate nursing students (n=16)	NurCaseAI
S16	Kroll et al. (2025)[23]	United States	Mixed-methods study	Simulation	Nurse educator and Advanced Practice Nurses (n=16)	Large Language Models (LLMs) GPT-4 GPT-4o
S18	Han et al. (2025)[24]	South Korea	Quasi-experimental study	Online	Undergraduate nursing students (n=61)	AI Chatbot (LandBot)
S19	Han et al. (2025)[10]	South Korea	Randomized controlled trial	Online	Nursing students (n=60)	AI chatbot (LandBot)
S20	Fung et al. (2025)[25]	Hong Kong	Cross-over randomized controlled trial	Simulation	Undergraduate nursing students (n=44)	GenAI patient simulation system
S21	Park&Kim (2025)[26]	South Korea	Quasi-experimental study	Simulation	Undergraduate nursing students (n=72)	AI tutor (Danbee platform)

Early Version - Uncorrected

Study ID	Authors (Year)	Country/Region	Study design	Setting	Participants (n)	AI tool/Model
S22	Wolf&Wolf (2025)[27]	United States	Quasi-experimental study	Online	Master ’s-level nursing students (n=64)	IntelliMetric® AES

Early Version - Uncorrected

Table 2 Application characteristics of AI-generated feedback in nursing documentation and clinical reasoning training

Study ID	AI Technique/ System	Clinical Scenario	Type of Documentation	Feedback Output	Timing	Target competency	Adaptivity
S05	SmartENR AI (OpenAI ChatGPT 4.0 API)	110 virtual patient scenarios (e.g., pressure ulcer, post-stroke)	NANDA, SOAPIE, Focus DAR, Narrative formats	AI-generated recommended nursing records	Real-time	Nursing diagnosis & documentation efficiency	Unspecified
S06	4A program (ChatGPT-4o API, STT, NLP)	Maternal and child health nursing (3 case studies)	Audio responses (transcribed to text)	Semantic similarity scoring vs. standard answers	Post-task	Cognitive abilities (Critical thinking, reasoning)	Unspecified
S07	Eval-Bot v1 (OpenAI- GPT-4.0)	Hospital emergency during a power outage	Chatbot dialogue/Crisis response prompts	Structured debriefing (strengths, tips, summary)	Post-task	Clinical decision-making & emergency planning	Specified: Role-based branching and experience level tailoring

Early Version - Uncorrected

Study ID	AI Technique/ System	Clinical Scenario	Type of Documentation	Feedback Output	Timing	Target competency	Adaptivity
S08	Virti AI platform (Generative AI/Virtual Human)	Psychiatric nursing (Suicide risk: MDD, Schizophrenia, etc.)	Reflection journals & dialogue logs	Quantitative scoring and expert qualitative insights	Real-time and post-task	Emotional Intelligence & Crisis Management	Specified: Student-led dynamic interactions
S09	AI-enhanced Mixed Reality (MR) training platform (built with MindMeld)	Physiological deterioration (Hypotension & Hypoxia)	Verbal queries (voice-to-text) and behavioral actions	Real-time verbal/gestural patient responses & SAGAT questions	Real-time and post-task	Clinical reasoning & situational awareness	Specified: Real-time contextual response to inputs
S10	Artificial Intelligence Integrated Nursing Process (AINT-P) training program	Clinical scenarios for nursing process	Nursing diagnoses, goals, and intervention plans	Feedback on diagnosis and intervention accuracy	Real-time	Problem-solving skills & nursing process competency	Unspecified

Early Version - Uncorrected

Study ID	AI Technique/ System	Clinical Scenario	Type of Documentation	Feedback Output	Timing	Target competency	Adaptivity
S13	ChatGPT 4.0 (Question-centered model)	Ileus clinical reasoning scenario	Model-based worksheets (staged prompts)	AI-generated clinical reasoning and step-by-step responses	Real-time	Creative thinking & clinical reasoning	Specified: Structured integration within question phases
S14	NurCaseAI RAG (Retrieval-Augmented Generation)	Thyroid perioperative nursing	Interactive Q&A and comprehensive reports	Targeted feedback, score summaries, and learning suggestions	Real-time	Clinical reasoning & critical thinking	Specified: Dynamically adjusts learning paths based on input
S16	Dual-agent LLM framework (GPT-4, GPT-4o)	SBIRT training (Screening, Brief Intervention, Referral to Treatment) for Substance Use Disorder (SUD)	Conversation logs (text or microphone)	Immediate assessment based on GITCS criteria	Post-task	Therapeutic communication & SBIRT skills	Specified: LLM adapts patient openness based on communication

Early Version - Uncorrected

Study ID	AI Technique/ System	Clinical Scenario	Type of Documentation	Feedback Output	Timing	Target competency	Adaptivity
S18	AI-based “talking bot” developed via LandBot	Electronic Fetal Monitoring (EFM)	Quizzes, MCQ, and open-ended questions	Immediate corrective feedback and content summaries	Real-time and post-task	Clinical reasoning (EFM skills) & self-directed learning	Specified: Customized content/re-study path for incorrect answers
S19	AI chatbot (LandBot.io, NLP, NLU)	Mechanical ventilation nursing	Interaction dialogues, formative quizzes, clicks	Personalized feedback, encouraging messages, and multimedia	Real-time and post-task	Clinical reasoning, knowledge, and self-confidence	Specified: Personalized interactions and optional additional study
S20	Generative AI (LLM), Patient Simulation	Acute pneumonia & acute appendicitis cases	Interactive dialogue (history-taking/physical exams)	Constructive expert feedback via "GenAI debriefing"	Real-time and post-task	Clinical competence, cultural awareness, and AI readiness	Specified: Tailored learning pathways and customized feedback

Early Version - Uncorrected

Study ID	AI Technique/ System	Clinical Scenario	Type of Documentation	Feedback Output	Timing	Target competency	Adaptivity
S21	AI tutor chatbot (interactive environment)	Labor nursing scenarios (NST, Leopold's, etc.)	Interaction queries and worksheets	Individualized feedback, prompts, or hints	Real-time and post-task	Clinical performance competency & digital literacy	Specified: Personalized learning tailored to academic level
S22	IntelliMetric® (Automated Essay Scoring - AES)	Scholarly writing workshop for master's level	Nursing documentation	Quantitative holistic and subscale writing scores	Real-time and post-task	Scholarly writing performance & critical thinking	Specified: Results used for "flexible scale of help" differentiation

Early Version - Uncorrected

Table 3 Validation approaches and trustworthiness evidence for AI-generated feedback systems (0–2 scoring per domain)

Study ID	Expert Review	Human Agreement	Guideline/ Rubric	Reliability/ Consistency	Robustness across cases	Safety/ Risk	Strong Validation Domains	Core Validation (Yes/No)
S05	2	1	2	2	2	1	4	Yes
S06	2	2	2	2	1	1	4	Yes
S07	2	2	2	1	1	2	4	Yes
S08	2	1	2	1	2	1	3	Yes
S09	2	1	2	2	2	2	5	Yes
S10	1	1	1	2	1	1	1	Yes
S13	2	2	2	2	1	2	5	Yes
S14	1	2	2	2	2	2	5	Yes
S16	2	2	2	1	1	1	3	Yes
S18	2	1	2	2	1	1	3	Yes
S19	2	1	2	2	2	1	4	Yes

Early Version - Uncorrected

Study ID	Expert Review	Human Agreement	Guideline/ Rubric	Reliability/ Consistency	Robustness across cases	Safety/ Risk	Strong Validation Domains	Core Validation (Yes/No)
S20	1	2	2	2	2	1	4	Yes
S21	2	1	2	2	2	1	4	Yes
S22	2	2	2	2	2	1	5	Yes

Domain scores were assigned as follows: 0 = not reported; 1 = reported with limited detail; 2 = reported with explicit methods or metrics.
Strong domains indicate validation domains with explicit evidence (score = 2).
Core validation is reported descriptively; a "Yes" means the presence of at least one domain with explicit validation evidence and does not imply overall validation quality.

Early Version - Uncorrected