BMJ Open

Pedagogical strategies for enhancing critical thinking in nursing students using large language models: a mixed-methods systematic review protocol

Lucy Oh ,[1] Gael Mearns,[1] Rebecca Mowat[2]

To cite: Oh L, Mearns G, Mowat R. Pedagogical strategies for enhancing critical thinking in nursing students using large language models: a mixed-methods systematic review protocol. *BMJ Open* 2026;**16**:e119386. doi:10.1136/bmjopen-2026-119386

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2026-119386).

Received 09 March 2026
Accepted 12 June 2026

[1]School of Nursing, Faculty of Health and Environmental Sciences, Auckland University of Technology, Auckland, New Zealand
[2]School of Health, University of the Sunshine Coast, Sippy Downs, Queensland, Australia

Correspondence to
Lucy Oh; lucy.oh@aut.ac.nz

ABSTRACT

Introduction Large language models (LLMs) are rapidly normalising in nursing education. However, evidence regarding their impact on critical thinking—a cornerstone of safe clinical practice—remains inconsistent. Current syntheses have largely focused on student satisfaction and general utility, leaving the specific instructional mechanisms that cultivate this competency unexplored. This mixed-methods systematic review aims to identify effective pedagogical strategies for fostering critical thinking, providing evidence-based guidance to safeguard clinical competency in an artificial intelligence-integrated workforce.

Methods and analysis This review will follow a convergent segregated mixed-methods design in accordance with the Joanna Briggs Institute (JBI) methodology. A systematic search will be conducted across CINAHL, ERIC, MEDLINE, PsycINFO and Scopus supplemented by grey literature sources (ProQuest Dissertations, MedRxiv). The search will be limited to English-language studies published from January 2023 to June 2026. Eligible studies will include undergraduate nursing students exposed to structured LLM-based pedagogical interventions. Methodological quality will be appraised using design-specific JBI Critical Appraisal Checklists. Outcomes will be classified using a modified Kirkpatrick framework to distinguish objective competence from subjective confidence. Quantitative data will be synthesised narratively following Synthesis Without Meta-analysis guidelines, and qualitative data thematically. Streams will be integrated using a Joint Display Analysis to generate meta-inferences on effective instructional design. Finally, confidence in the cumulative evidence will be assessed using Grading of Recommendations Assessment, Development and Evaluation (GRADE) and GRADE-Confidence in the Evidence from Reviews of Qualitative research.

Ethics and dissemination As a systematic review of existing literature, this study does not require ethical approval. Findings will be disseminated through a peer-reviewed publication, conference presentations and a visual guide for educators.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This review employs a modified Kirkpatrick framework to distinguish between 'objective competence' and 'subjective confidence' preventing the methodological conflation of student satisfaction with actual competency gains.
⇒ The convergent segregated mixed-methods design integrates quantitative effectiveness data with qualitative student experiences, enabling the review to explain not just if an intervention works, but how and why specific pedagogical strategies foster critical thinking.
⇒ To distinguish robust evidence from preliminary exploratory studies common in this nascent field, the review uses domain-specific Joanna Briggs Institute appraisal tools to inform a stratified Grading of Recommendations Assessment, Development and Evaluation (GRADE) and GRADE-Confidence in the Evidence from Reviews of Qualitative research assessment of certainty.
⇒ Statistical meta-analysis is unlikely to be feasible due to the anticipated heterogeneity of pedagogical interventions and outcome measures, necessitating a narrative synthesis approach.
⇒ The restriction to English-language publications may introduce a geographical bias, potentially excluding relevant non-Western perspectives on artificial intelligence integration in nursing curricula.

INTRODUCTION

Large language models (LLMs) are a subset of text-based Generative Artificial Intelligence (GenAI) systems trained on large-scale language data to produce contextually responsive, natural-language outputs from user prompts.[1] The integration of these models in nursing education is moving rapidly from novelty to normalisation.[1 2] This transition is evident in students' self-directed learning practices, where LLMs are employed to synthesise clinical information and expedite academic workflows—such as drafting care plans, soliciting feedback and resolving complex case studies.[1–3] Current scholarship has effectively mapped the topography of adoption, identifying the widespread prevalence of these GenAI tools and their perceived

utility for enhancing educational efficiency.[1 3 4] Given this ubiquity, the urgent imperative is moving beyond the question of whether LLMs should be used, to a rigorous examination of how their integration shapes the professional development of future nurses.

Central to this development is critical thinking. Defined by the American Philosophical Association's seminal Delphi Report as 'purposeful, self-regulatory judgement' comprising both cognitive skills and affective dispositions,[5] critical thinking serves as the essential driver for clinical reasoning and the ultimate outcome of clinical judgement.[6] Unlike simple knowledge retrieval, this multifaceted competency requires the ability to plan, monitor and evaluate one's own thinking—metacognitive functions that are indispensable for navigating clinical ambiguity and ensuring patient safety.[6]

However, emerging evidence indicates that the impact of LLMs on critical thinking is highly variable. While recent syntheses confirm that LLM-based interventions generally improve learner satisfaction and personalised knowledge acquisition,[3 4] findings regarding higher-order thinking remain inconsistent.[3 7] Empirical studies reveal a sharp divergence in outcomes: while some structured pedagogical approaches appear to foster deep analytical skills,[8–10] others fail to yield measurable competency gains or even inadvertently encourage passivity.[11–13]

This inconsistency suggests that the cognitive impact of LLMs is not intrinsic to the tool, but dependent on the pedagogical design. A preliminary search of PROSPERO, MEDLINE and CINAHL revealed no current or ongoing review addressing this specific evidence gap. While existing syntheses identified in this search have confirmed these mixed outcomes, they have largely focused on general utility or subjective satisfaction, leaving the specific mechanisms that drive critical thinking outcomes unexplored.[3 4 7] What is now required is a systematic understanding of the 'active ingredients'—the specific instructional strategies—that constitute effective pedagogical design, providing the evidence-based guidance necessary to safeguard clinical competency in an AI-integrated workforce.

AIMS

The overarching aim of this review is to synthesise the current evidence on LLMs in nursing education to identify effective pedagogical strategies for fostering critical thinking. Employing a mixed-methods convergent segregated approach,[14] this review specifically seeks to identify and characterise the instructional design features (eg, scaffolding, prompt frameworks) associated with the successful development of this core competency. The ultimate objective is to provide an explanatory synthesis of what strategies work (effectiveness) and why they work (perception), thereby informing evidence-based guidance for safeguarding professional competency in an AI-enabled curriculum.

To achieve this, the review will address two distinct research questions:

1. **Quantitative question:** What is the effectiveness of pedagogical strategies documented in the literature for enhancing critical thinking among nursing students using LLMs?
2. **Qualitative question:** How do nursing students perceive or describe the influence of these pedagogical strategies on the development of their critical thinking?

METHODS AND ANALYSIS

Study design

This protocol has been developed in accordance with the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) guidelines.[15] This review will follow a convergent segregated mixed-methods design in accordance with the Joanna Briggs Institute (JBI) methodology for mixed-methods systematic reviews.[14] This design was selected as the most appropriate rigorous framework for addressing the nascent and heterogeneous nature of the literature on LLMs.

While quantitative data can quantify whether an intervention works, and qualitative data explores how it is experienced, neither can provide a comprehensive understanding of impact across different contexts in isolation. Therefore, a segregated approach is used to independently synthesise these streams while preserving their methodological integrity. This separation enables the final integration to generate explanatory meta-inferences, using student perceptions to contextualise why effectiveness outcomes vary between specific pedagogical designs.

The review is planned to commence in June 2026 and is expected to be completed by March 2027.

Eligibility criteria

While the overarching aim of this review focuses on critical thinking, primary empirical literature in nursing education frequently uses this term interchangeably with 'clinical reasoning' and 'clinical judgement' or conceptualises them as a developmental continuum.[6] To ensure maximum sensitivity and capture all relevant data regarding this core competency, this review operationally groups critical thinking, clinical reasoning and clinical judgement as collective indicators of the outcome of interest.

The specific eligibility criteria are structured using two distinct frameworks to align with our segregated design: PICO (Population, Interventions, Comparators, Outcomes) for the quantitative review question (defined in table 1), and PICo (Population, Phenomenon of Interest, Context) for the qualitative review question (defined in table 2). The detailed inclusion and exclusion criteria derived from these frameworks are provided in online supplemental file 1.

Table 1 PICO framework

Element	Description
P (Population)	Undergraduate or preregistration nursing students.
I (Intervention)	Any defined pedagogical strategy, activity or intervention designed to foster critical thinking, clinical reasoning or clinical judgement during the use of large language models.
C (Comparator)	Any comparator (eg, standard curriculum, alternative pedagogical strategies or no intervention) or no comparator (eg, single-group pre–post designs).
O (Outcome)	Any measurable outcomes related to critical thinking, clinical reasoning or clinical judgement.

Search sources

A systematic search will be conducted in the following electronic databases:

- CINAHL (Cumulative Index to Nursing and Allied Health Literature) via EBSCOhost
- ERIC (Education Resources Information Center) via Ovid
- MEDLINE via EBSCOhost
- PsycINFO via Ovid
- Scopus

To identify relevant unpublished empirical research, the following grey literature sources will also be searched:

- Dissertations & Theses: ProQuest Dissertations & Theses Global
- Pre-print Servers: medRxiv, Preprints.org
- Trial Registries: ClinicalTrials.gov, World Health Organization International Clinical Trials Registry Platform (WHO ICTRP)

Search strategy

A comprehensive search strategy will be developed in collaboration with a health sciences information specialist. The strategy will be systematically built by combining two different search techniques to ensure maximum sensitivity and precision, in line with the PRISMA-S (Search Reporting) extension.[16]

The strategy will be developed iteratively. It will combine three core concepts: Population (P), the LLM Intervention (I), and the Outcome (O)/Phenomenon of Interest. The 'Pedagogy' component of the intervention will not be used as a search block; it will be applied as a screening criterion during study selection (per eligibility criteria). This is a deliberate methodological choice to maximise sensitivity and avoid missing relevant studies in this emerging field, where pedagogical terms may not be used consistently in abstracts.

Table 2 PICo framework

Element	Description
P (Population)	Undergraduate or preregistration nursing students.
I (Phenomenon of Interest)	Students' experiences, perceptions, descriptions or perceived mechanisms related to the influence of pedagogical strategies on their critical thinking, clinical reasoning or clinical judgement when using large language models.
Co (Context)	Higher education nursing programmes or any post-secondary educational setting for nursing.

Prior to execution, the draft strategy (eg, for MEDLINE) will be peer-reviewed by a second information specialist using the PRESS (Peer Review of Electronic Search Strategies) checklist.[17] The final, as-executed search strategy for all databases will be documented and provided in a supplementary appendix for the final review.

The search will be built using the specific subject headings and keywords detailed in table 3:

- Controlled Vocabulary (Subject Headings): For the stable concepts of this review (Population and Outcome), a comprehensive set of official subject headings (eg, CINAHL Headings, MeSH, PsycINFO Thesaurus) will be used. This ensures studies are captured based on their indexed topic, regardless of the specific terminology used by the authors.
- Keywords (Free-Text): For the emerging concept (the Intervention), a rigorous keyword-only search will be used. This is essential as the terminology for LLMs (eg, "ChatGPT", "LLM") is new, and stable subject headings do not yet exist. These keywords will be searched in the title and abstract fields. While our review focuses on LLMs, broad terms (eg, 'Generative AI') will be included in the search strategy to ensure sensitivity, as authors frequently use these umbrella terms in titles even when investigating specific text-based models.
- Grey literature strategy: A more sensitive, keyword-only strategy will be adapted for the grey literature sources, which do not support controlled vocabularies. This strategy will primarily combine the core keyword concepts for the Intervention (eg, "ChatGPT") and the Population (eg, "nursing student*").
- Handsearching: The reference lists of all included studies and any identified systematic or scoping reviews will be handsearched to identify any potentially eligible primary studies that were not captured by the electronic search.

A complete draft search strategy for MEDLINE (via EBSCOhost), including all subject headings and keywords, is provided in online supplemental file 2.

Study selection and data extraction

Record management

Search results will be imported into Covidence (Veritas Health Innovation, Melbourne, Australia), a web-based software platform designed for streamlining systematic reviews, for deduplication and screening.

Selection process

The selection process will proceed in three stages:

Table 3 Keywords search strategy

Concept (from PICO/PICo)	Subject headings (controlled vocabulary)	Keywords (free-text)
P (Population)	CINAHL Headings: Students, Nursing Students, Nursing, Baccalaureate Education, Nursing, Baccalaureate Education, Nursing, Diploma MeSH (MEDLINE): Students, Nursing Education, Nursing, Baccalaureate Education, Nursing, Diploma Programs PsycINFO Thesaurus: Nursing Students Nursing Education (Undergraduate Students AND Nursing)	("nursing student*" OR "student nurse*" OR "undergraduate nurse*" OR "undergraduate nursing" OR "nursing undergraduate*" OR "baccalaureate nursing" OR "BSN" OR "BNurs*" OR "pre-registration nurs*" OR "preregistration nurs*" OR "pre-reg nurs*" OR "diploma nurs*" OR "student in nursing" OR "undergraduate in nursing")
I (Intervention)	[Not used due to emerging terminology])	("generative AI" OR "generative artificial intelligence" OR "large language model*" OR "LLM*" OR "text-generating AI" OR "ChatGPT*" OR "GPT*" OR "Gemini" OR "Claude" OR "Copilot")
O (Outcome)/I (Phenomenon of Interest)	CINAHL Headings: Critical Thinking Clinical Reasoning Clinical Judgment Problem Solving Decision Making Metacognition Clinical Competence MeSH (MEDLINE): Thinking Judgment Clinical Decision-Making Problem Solving Metacognition Clinical Competence PsycINFO Thesaurus: Critical Thinking Problem Solving Decision Making Metacognition Clinical Reasoning	("critical thinking" OR "higher-order thinking" OR "HOTS" OR "clinical reasoning" OR "clinical judgment" OR "clinical judgement" OR "clinical decision-making" OR "decision making" OR "problem solving" OR "diagnostic reasoning" OR "analytical skill*" OR "evidence-based practice" OR "evidence based practice" OR ethical reasoning OR metacognition OR "self-regulation" OR "critical thinking disposition*")

1. Piloting: Prior to formal screening, the review team will jointly screen a pilot sample (eg, 50–100 abstracts) to calibrate the application of the eligibility criteria. The team will meet to discuss all conflicts, refine the decision rules (particularly for the 'Intervention' definition in online supplemental file 1), and ensure a high level of inter-rater reliability before proceeding.
2. Screening: Following the pilot, two reviewers will independently screen all remaining titles and abstracts against the eligibility criteria.
3. Full-text review: The same two reviewers will independently assess the full text of all potentially relevant articles.

Any conflicts at either stage will be resolved through discussion or, if necessary, by a third reviewer. A PRISMA flow diagram will document the process.

Data collection process

Two reviewers will independently extract data using a pre-piloted data extraction form. The specific variables to be extracted, aligning with our mixed-methods convergent segregated design, are detailed in table 4.

Table 4 Data extraction fields

Data category	Framework	Variables to be extracted
1. General characteristics	Bibliographic and design	► Bibliographic: Author(s), year, country of origin, funding source. ► Study design: Stated aim(s), specific design (eg, RCT, phenomenology).
2. Intervention description (all studies)	TIDieR [18]	► What: Specific LLM tool, prompts, tasks, rubrics used. ► How/Who: Delivery mode (eg, simulation) and facilitator role. ► Dose: Duration, frequency, total exposure.
3. Quantitative data	PICO	► P (Population): Quantitative sample size, participant characteristics (eg, year level, setting). ► I (Intervention): See row 2, "Intervention description (all studies)". ► C (Comparator): Details of control/alternative intervention. ► O (Outcomes): All measured outcomes related to critical thinking, clinical reasoning or clinical judgement (including both objective scores and subjective ratings): Instrument names (eg, HSRT, CCTDI). Psychometric properties reported? (Yes/No/Citation of validation study). Summary statistics (means, SDs, p values, effect sizes).
4. Qualitative data	PICo	► P (Population): Qualitative sample size, participant characteristics. ► I (Phenomenon of Interest): Student experiences, perceptions or descriptions regarding the intervention, including perceived barriers, facilitators and mechanisms of learning. ► Co (Context): Educational setting. ► Findings: Direct participant quotes and author-derived themes.
5. Mixed-methods integration (If applicable)	Joint display context	► Primary Author Integrations: Explicitly stated meta-inferences, explanatory linkages or triangulated findings that connect the quantitative and qualitative strands within the primary study.

CCTDI, California Critical Thinking Disposition Inventory; HSRT, Health Sciences Reasoning Test; LLM, large language model; RCT, randomised controlled trial; TIDieR, Template for Intervention Description and Replication.

Data handling and management

- **Mixed-methods extraction and segregation:** In accordance with the JBI convergent segregated approach, quantitative and qualitative data from mixed-methods studies will be extracted and channelled into their respective independent synthesis streams.[14] To ensure the philosophical integrity of primary mixed-methods research is maintained, reviewers will also explicitly extract any explanatory linkages, triangulated findings or meta-inferences generated by the primary authors. These original linkages will be preserved as contextual data to inform the final Joint Display Analysis. In instances where a primary study's data is so deeply integrated that clean separation into quantitative and qualitative streams is not possible without losing meaning, the research team will review the study and reach a consensus on how best to extract and manage the findings to preserve the authors' original intent.
- **Outcome sorting:** All reported outcomes will be extracted initially. To ensure consistency, the classification of these outcomes into Modified Kirkpatrick levels (detailed in the 'Outcomes and Prioritisation' section below) will be performed as a postextraction analysis step:
  - Numerical subjective data (eg, test scores, Likert scale scores) will be channelled into the quantitative synthesis.
  - Textual data (eg, interview narratives, open-text responses) will be channelled into the qualitative synthesis.
- **Missing data strategy:** If key data (eg, means, SD) are not reported in the text but are available in figures, we will use WebPlotDigitizer (Automeris, Pacifica, California, USA) to extract them. For crucial missing data, we will contact the corresponding authors a maximum of two times over a 4-week period. If unresolved, the study will be analysed using the available reported data, and the omission explicitly noted as a limitation.

Outcomes and prioritisation

The framework and justification

To systematically manage the anticipated heterogeneity in reported outcomes, this review uses an adapted version of the New World Kirkpatrick Model[18] as a classification framework. The Kirkpatrick model is widely recognised as the standard for evaluating training effectiveness, hierarchically categorising outcomes into four distinct levels: level 1 reaction (participant satisfaction), level 2 learning (acquisition of knowledge, skills or confidence), level 3 behaviour (application of learning in practice) and level 4 results (organisational impact). This structure is particularly well-suited to a mixed-methods design as it captures the full spectrum of educational impact, accommodating both subjective user experience and objective competency gains.

Rationale for adaptation

While the New World Kirkpatrick Model provides a comprehensive structure, its standard definition of level 2 (learning) aggregates diverse constructs including 'knowledge, skills, attitude, confidence and commitment'.[18] In the context of standard professional training, combining 'confidence' (the belief one can perform) with 'skill' (the ability to perform) is conventionally accepted, as this framework posits that confidence is a key indicator of learning.

However, in the nascent field of LLM-based education, emerging evidence suggests a potential disconnect between student confidence and actual competence.[3 7] Students may report high confidence (a subjective level 2 outcome) without possessing the requisite critical thinking skills (an objective level 2 outcome).

Therefore, strict adherence to the standard model—which would group a self-reported increase in confidence alongside a validated test score as simply 'level 2 learning'—risks obscuring the true pedagogical impact.

The modified approach

To mitigate this validity threat, the review employs the modified Kirkpatrick model illustrated in figure 1. As depicted, our protocol explicitly bifurcates level 2. We will rigorously distinguish between objective competence (levels 2A and 2B) and subjective perception (level 2C).

- **Objective competence outcomes:** Findings mapped to level 3 (behaviour) and levels 2A/2B (objective learning) are prioritised as high-utility evidence of pedagogical effectiveness. These represent measurable changes in student competence.
- **Subjective perception outcomes:** Findings mapped to level 2C (subjective learning) and level 1 (reaction) are treated as contextual evidence. Depending on the data type, these will be analysed either narratively (for numerical metrics) to contrast confidence with competence, or thematically (for textual narratives) to explain how and why an intervention works (see Data synthesis section).

(Note: Level 4 (results) of the original model[18] is excluded, as outcomes related to organisational impact or patient safety are considered outside the scope of this review).

The specific definitions, examples and synthesis prioritisation for each level are detailed in table 5.

Hierarchy of evidence and prioritisation strategy

To ensure the robustness of our findings, the evidence tiers defined in table 5 will be handled separately during synthesis. This prioritisation ensures that findings from high-quality evidence are not diluted by findings from instruments of unknown validity.

1. Primary evidence (high-utility evidence): Evidence from studies using controlled designs, including randomised controlled trials (RCTs) and quasi-experimental studies with a comparator, and reporting level 3 or level 2A outcomes will be treated as the

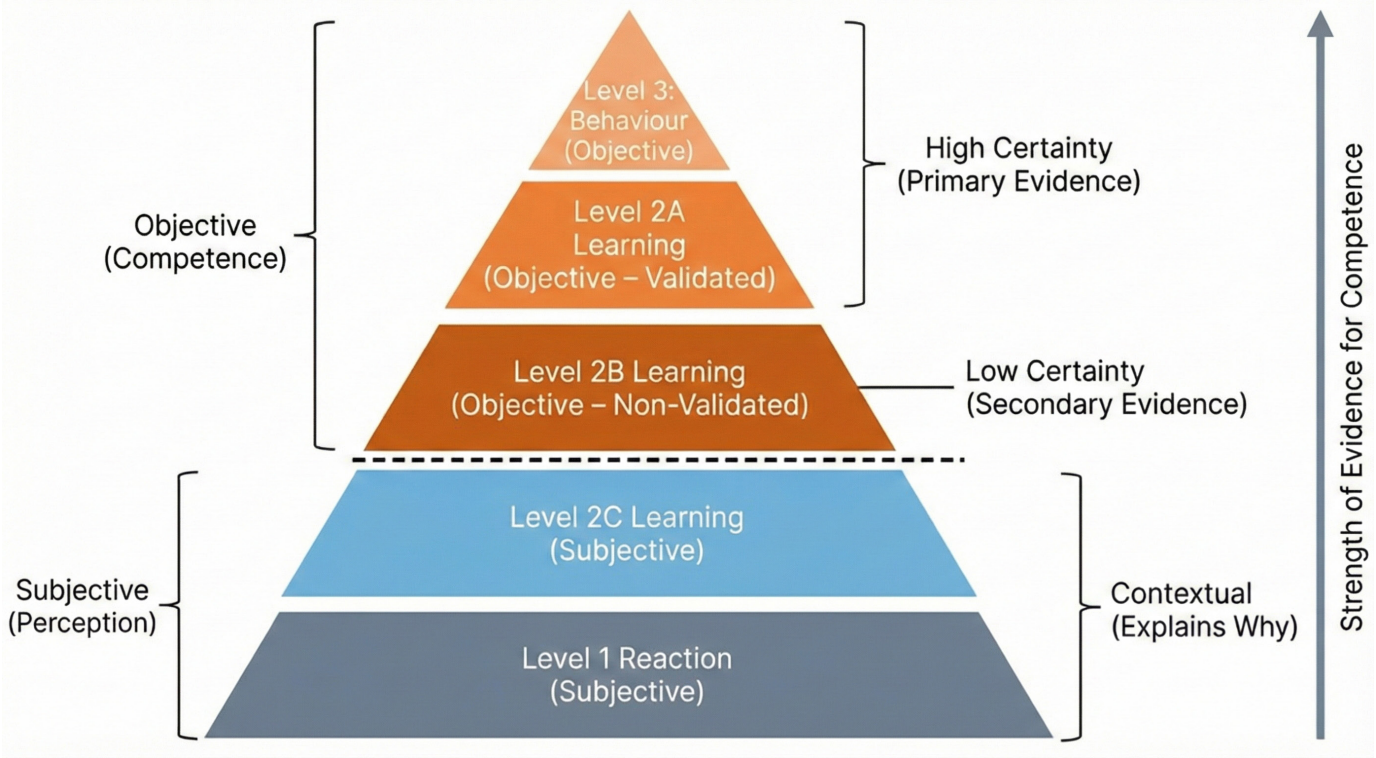

Figure 1 The modified Kirkpatrick framework for outcome classification. The model hierarchically prioritises evidence by distinguishing objective competence (levels 3, 2A, 2B) from subjective perception (levels 2C, 1). The dashed line indicates the bifurcation of level 2 (learning), specifically separating measured skills from self-reported confidence.

Table 5 Modified Kirkpatrick outcome classification matrix

Construct domain	Modified level	Definition
Objective (Competence)	Level 3 Behaviour (Objective)	Observable changes in the application of critical thinking or clinical reasoning in high-fidelity simulation or authentic clinical practice. Example: ► Directly observed performance in a clinical setting (eg, via a validated competency tool).
	Level 2A Learning (Objective–Validated)	Objective, quantifiable changes in critical thinking knowledge or skills as measured by established, psychometrically validated instruments. Example: · ► Health Sciences Reasoning Test. ► California Critical Thinking Skills Test.
	Level 2B Learning (Objective–Non-Validated)	Objective, quantifiable changes in the application of skills measured by bespoke, author-developed or non-validated tools (validity/reliability unknown). Example: · ► Author-developed multiple-choice quizzes or exams specific to the study. ► Bespoke/unvalidated rubrics.
Subjective (Perception)	Level 2C Learning (Subjective)	Student self-perceptions of their own growth in critical thinking, confidence or critical dispositions (eg, scepticism). Example: · ► Self-reported confidence or ability in clinical reasoning. ► Perceived knowledge gain. ► Changes in AI scepticism or reliance.
	Level 1 Reaction (Subjective)	Student perceptions of the pedagogical activity (barrier/facilitator). Example: · ► Course satisfaction surveys. ► Perceived utility of the AI tool for the assigned task.

AI, artificial intelligence.

primary evidence for answering subquestion 1 (effectiveness). These data will form the core of our primary narrative synthesis.
2. Secondary evidence (low-utility evidence): Evidence from studies using uncontrolled designs (eg, single-group pre–post) or reporting level 2B outcomes will be synthesised separately. Findings based solely on these sources will be explicitly designated as low-certainty evidence of effectiveness due to maturation threats and measurement bias.
3. Contextual evidence: Findings from studies reporting level 1 or level 2C outcomes will be used to provide explanatory context.
   - Numerical metrics (surveys): Will be reported narratively in the quantitative stream to identify confidence-competence gaps.
   - Textual narratives (interviews): Will be synthesised thematically to answer the qualitative subquestion (subquestion 2).
   - Note: Both types are explicitly designated as insufficient evidence for establishing pedagogical effectiveness.

Risk of bias in individual studies

To ensure strict methodological congruence with our chosen Convergent Segregated approach, the methodological quality of all included studies will be appraised using the suite of JBI Critical Appraisal Checklists.[14] Mixed-methods studies will be appraised by assessing their quantitative and qualitative components independently using the corresponding design-specific JBI tool.

Given the anticipated heterogeneity of study designs in this nascent pedagogical field, the JBI suite was selected to provide a unified, design-specific appraisal ecosystem. This choice balances two competing methodological needs: sensitivity and granularity. Unlike clinical tools such as Cochrane's RoB 2[19] and ROBINS-I,[20] the JBI checklists mitigate the risk of floor effects inherent in applying rigid clinical standards to educational interventions. Furthermore, unlike the streamlined Mixed Methods Appraisal Tool,[21] the JBI suite offers the domain-level granularity (eg, detailed assessments of allocation concealment and blinding) explicitly required to inform our planned GRADE (Grading of Recommendations Assessment, Development and Evaluation)[22] and GRADE-CERQual (Confidence in the Evidence from Reviews of Qualitative research)[23] certainty assessments.

The appraisal will be conducted by two reviewers independently, with disagreements resolved through consensus or a third reviewer. No studies will be excluded based on methodological quality; excluding studies in this nascent, fragmented field would result in a loss of valuable data.

Instead, the granular results of the JBI critical appraisals will be used to:

1. Describe the overall methodological quality of the included evidence.
2. Conduct a sensitivity analysis to distinguish between findings supported by robust evidence versus those at high risk of bias.
3. Inform the final assessment of confidence in the cumulative evidence using GRADE for quantitative findings[22] and GRADE-CERQual for qualitative findings.[23]
4. Designate unpublished studies (eg, conference abstracts) as 'unclear' quality where full appraisal is not possible, ensuring findings are synthesised with appropriate caution.

Data synthesis

As statistical pooling is not anticipated due to the heterogeneity of pedagogical strategies, our mixed-methods synthesis will follow a Convergent Segregated approach.[14] This will involve two distinct synthesis streams followed by a final integration, as illustrated in figure 2.

Quantitative synthesis

A narrative synthesis will be performed for all numerical data, guided by the framework from Popay *et al*[24] and adhering to the Synthesis Without Meta-analysis reporting guidelines.[25]

- Intervention taxonomy (grouping strategy): As a formal taxonomy for LLM-based pedagogy does not exist, we will develop one using an abductive approach. While all included studies must meet the deductive criteria for 'Intentional Design' (see online supplemental file 1), the specific pedagogical mechanisms (eg, adversarial prompting, scaffolded verification) remain undefined in the literature. Therefore, we will iteratively sub-classify the intervention descriptions, extracted using the Template for Intervention Description and Replication (TIDieR), into a coherent framework using qualitative content analysis[26]:
  - Coding: Two reviewers will independently code the intervention descriptions for core pedagogical components (eg, 'prompt-writing exercise', 'AI-as-tutor').[27]
  - Categorisation: Using an iterative process of comparing differences and similarities, these codes will be grouped into higher-level categories to build the taxonomy.
  - Consensus: Disagreements regarding classification will be resolved through consensus or arbitration by a third reviewer.
- Synthesis and prioritisation: Findings will be grouped by these taxonomy categories. To rigorously examine the relationship between competence and confidence, quantitative data will be synthesised in two streams:
  - Objective effectiveness (levels 2A/B and 3): Standardised test scores, rubrics and behavioural checklists. We will explicitly distinguish findings extracted via controlled studies (RCTs, quasi-experimental) from those extracted via uncontrolled studies (eg, single-group pre–post). Findings from uncontrolled studies will be treated as indicative only and

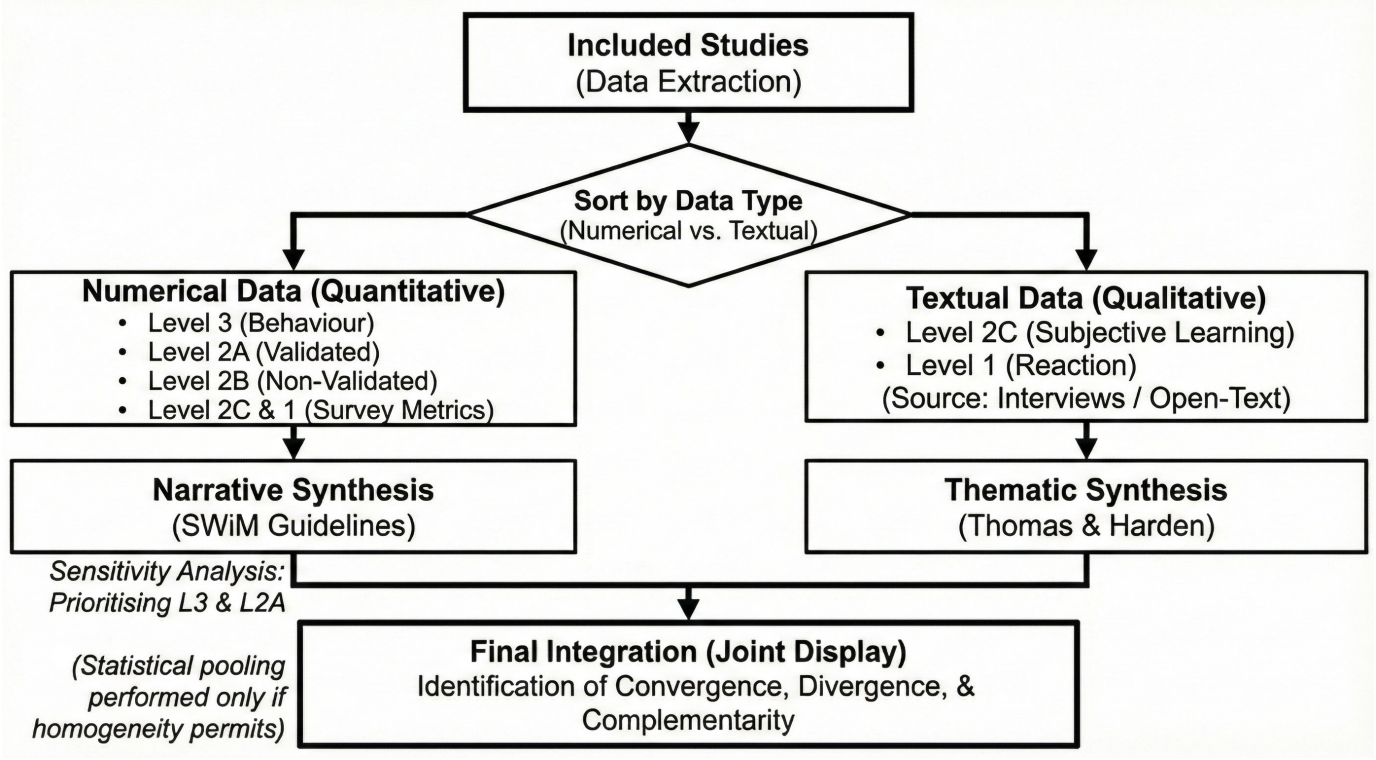

Figure 2 Convergent segregated synthesis schema. This diagram illustrates the separation of findings based on data type (numerical vs textual) and their distinct synthesis pathways (narrative vs thematic). The modified Kirkpatrick framework is applied within these streams to prioritise evidence prior to final integration via joint display analysis. SWiM, Synthesis Without Meta-analysis.

explicitly designated as insufficient evidence for establishing pedagogical effectiveness.
- Subjective metrics (levels 1 and 2C): Quantitative survey data (eg, Likert scales) measuring learner satisfaction and self-reported confidence. While these metrics will be reported narratively, they will be explicitly contrasted against the objective level 2A/B outcomes within the same studies to identify potential confidence-competence gaps before the final mixed-methods integration.

- Meta-analysis contingency: A statistical meta-analysis will only be performed in the unlikely event that a subset of studies is identified as sufficiently homogenous in terms of population, intervention, and outcome.
- Investigation of heterogeneity: We will assess the robustness of our findings through:
  - Sensitivity analysis: Comparing the primary synthesis findings against a dataset excluding studies appraised as at 'high risk of bias' according to the JBI checklists to ensure conclusions are not driven by biased evidence.
  - Subgroup analysis: Exploring potential differences in effect based on intervention type, population and outcome (if sufficient data exists).

All effectiveness outcomes (level 3, level 2A/B) will be entered into the narrative synthesis, but findings relying solely on level 2B data will be distinguished via the planned sensitivity analysis.

Qualitative synthesis

A thematic synthesis will be performed strictly for qualitative textual data (eg, interviews, open-ended survey responses), following the methodology of Thomas and Harden.[28] This involves a three-stage process: line-by-line coding of findings, the development of descriptive themes and the generation of analytical themes. The synthesis will be managed using qualitative data analysis software NVivo (Lumivero, Denver, Colorado, USA).

Integration of syntheses

Finally, the findings from the quantitative narrative synthesis (the 'what works') and the qualitative thematic synthesis (the 'why/how it works') will be integrated using a Joint Display Analysis approach[29]:

1. **Joint display construction:** The two streams of findings will be juxtaposed in a joint display table, aligning quantitative effectiveness data with qualitative perception themes for corresponding interventions. Concurrently, the explanatory linkages and meta-inferences extracted directly from primary mixed-methods studies will be integrated into this display to contextualise the findings and preserve the original authors' triangulated conclusions.

2. **Joint display analysis:** We will systematically compare the streams to identify patterns where specific qualitative findings (eg, student perceptions of usability) explain the presence, absence or variability of quantitative effects.
3. **Derivation of meta-inferences:** Based on this analysis, we will explicitly report on areas of convergence (mutual confirmation), divergence (contradiction) or complementarity (expansion) to characterise the specific pedagogical design features that facilitate critical thinking.

To ensure this integration is systematic, we define these relationships as follows:

- **Convergence (confirmation):** This will occur when qualitative findings from the thematic synthesis align with and support quantitative findings.
  - For example, qualitative themes on 'high student engagement' are matched with quantitative findings of 'high effectiveness scores' (levels 2A/2B or level 3) for the same intervention type.
- **Divergence (contradiction):** This will occur when qualitative and quantitative findings are in opposition.
  - For example, qualitative themes report 'high satisfaction' (level 1) but quantitative findings show 'no learning gains' (levels 2A/2B or level 3) for that intervention.
- **Complementarity (expansion/explanation):** This will occur when one synthesis explains or expands on the other.
  - For example, qualitative themes of 'poor instructions' or 'technical difficulties' (level 1) help to explain why an intervention showed no effect in the quantitative data.

Meta-bias(es)

We will assess outcome reporting bias (within-study bias) during the risk of bias appraisal by comparing the methods sections of included studies to their reported results to identify selectively omitted outcomes. Where available, we will also compare published results against original study protocols.

To mitigate publication bias (across-study bias), we have designed a comprehensive grey literature search strategy to identify unpublished findings (eg, dissertations, preprints) that may report null or negative results. As statistical assessment (eg, funnel plots) is not appropriate for this narrative synthesis, the potential impact of publication bias on the cumulative evidence will be discussed narratively.

Confidence in cumulative evidence

The overall confidence in the cumulative evidence will be assessed in a two-stage process to inform the final, integrated findings.

Stage 1: confidence in separate streams

- **Quantitative evidence:** The strength of the narrative synthesis evidence will be assessed using the GRADE framework.[22] We anticipate that the overall confidence in this stream will be 'low' or 'very low', given the necessary inclusion of non-randomised and single-group study designs to capture evidence from this nascent field.
- **Qualitative evidence:** Confidence in the thematic synthesis findings will be assessed using the GRADE-CERQual framework.[23]

Stage 2: confidence in integrated findings

Following the separate assessments, a final confidence rating (high, moderate, low or very low) will be assigned to each integrated finding (meta-inference). This rating will be a transparent, interpretive judgement guided by:

1. Input quality: The individual ratings from the underlying GRADE and CERQual assessments.
2. Relationship: The degree of convergence, complementarity or divergence between the streams.
3. Overall robustness: The quantity and consistency of the supporting data.

Decision matrix logic: We will employ a defined logic to assign final ratings. For example, an integrated finding supported by 'low' confidence quantitative evidence (effectiveness) may be upgraded to 'moderate' overall confidence if it is strongly corroborated by 'high' confidence qualitative evidence (perception) that explains the mechanism of action. The rationale for every final judgement, including any upgrades or downgrades, will be explicitly documented.

Patient and public involvement

Patients and the public were not involved in the design of this protocol, and will not be involved in the conduct, reporting or dissemination of the review.

ETHICS AND DISSEMINATION

Ethics statements

As this study is a systematic review of existing published literature, it does not involve primary data collection from human participants; therefore, formal ethical approval is not required. The review will be conducted in strict accordance with the JBI Manual for Evidence Synthesis and reported following the PRISMA 2020 to ensure methodological transparency and reproducibility. The protocol has been registered with the Open Science Framework (https://osf.io/d587j) to ensure accountability and prevent duplication of effort. Any important protocol amendments made during the conduct of the review will be documented, date-stamped and justified on the OSF registry and explicitly reported in the final manuscript.

Dissemination plan

The findings of this review will be disseminated through publication in a high-impact, peer-reviewed nursing education or digital health journal. Additionally, results will be presented at key international conferences to engage directly with the scholarship of teaching and learning community. To maximise accessibility for educators and policymakers, a summary of the key evidence-informed pedagogical strategies will be made available as a visual infographic and distributed via professional nursing networks.

Implications for practice and policy

This review responds to a critical transitional moment in the health workforce. As LLMs move from novelty to ubiquity, the primary risk to the nursing profession is no longer 'plagiarism' but cognitive offloading—the potential for the emerging workforce to defer clinical judgement to an algorithm. By systematically identifying pedagogical strategies that preserve and enhance critical thinking, this review will have three distinct impacts:

- **Workforce safety and clinical risk mitigation:** This review aims to mitigate the confidence-competence gap (the divergence between level 2C subjective confidence and level 2A objective competence). By identifying pedagogical strategies that genuinely foster objective competency rather than just passive reliance, this research directly contributes to patient safety by ensuring the future workforce is equipped to navigate clinical ambiguity and resist automation bias.
- **Evidence-based curriculum design:** Currently, AI integration in nursing curricula is largely experimental and fragmented. This review will provide educators with a taxonomy of 'active ingredients'—specific, replicable instructional designs (eg, adversarial prompting, scaffolded verification)—that have been empirically shown to trigger higher-order thinking, moving curriculum design from trial-and-error to evidence-based practice.
- **Governance and professional standards:** As health services and accrediting bodies seek to manage the risks of LLMs, this review will provide the evidence base to define what 'safe AI usage' looks like in practice. It moves the policy conversation beyond simple usage bans to the definition of robust governance frameworks, ensuring that digital literacy standards align with the realities of safe clinical judgement in an AI-integrated health system.

Contributors LO conceived the study, designed the protocol, created the visualisation models and drafted the original manuscript. GM and RM provided supervision, contributed to the refinement of the methodology and critically reviewed and edited the manuscript. All authors approved the submitted version of the manuscript and agree to be accountable for the work. LO is the guarantor for this review. Grammarly AI was used only for minor grammar and language polishing.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

Lucy Oh https://orcid.org/0009-0005-0611-2171

REFERENCES

1 Amankwaa I, Ekpor E, Cudjoe D, *et al*. Patterns, advances, and gaps in using ChatGPT and similar technologies in nursing education: A PAGER scoping review. *Nurse Educ Today* 2025;153:106822.
2 Yeh SN, Siah C-JR. Students' attitudes toward LLMs and its association with metacognitive abilities: A cross-sectional study. *Nurse Educ Pract* 2025;88:104567.
3 Ma J, Wen J, Qiu Y, *et al*. The role of artificial intelligence in shaping nursing education: A comprehensive systematic review. *Nurse Educ Pract* 2025;84:104345.
4 Labrague LJ, Sabei SA. Integration of AI-Powered Chatbots in Nursing Education: A Scoping Review of Their Utilization, Outcomes, and Challenges. *Teaching and Learning in Nursing* 2025;20.
5 Facione PA. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report); Millbrae (CA) California Academic Press; 1990.
6 Tanner CA. Thinking like a nurse: a research-based model of clinical judgment in nursing. *J Nurs Educ* 2006;45:204–11.
7 Liu Y-M, Chou C-C, Jaing T-H, *et al*. Generative AI for clinical reasoning: A scoping review. *Teaching and Learning in Nursing* 2026;21:e305–12.
8 Hwang G-J, Cheng P-Y, Chang C-Y. Facilitating students' critical thinking, metacognition and problem-solving tendencies in geriatric nursing class: A mixed-method study. *Nurse Educ Pract* 2025;83:104266.
9 Chang CY, Yang CL, Jen HJ, *et al*. Facilitating nursing and health education by incorporating ChatGPT into learning designs. *Educ Technol Soc* 2024;27:215–30.
10 Yin J, Hao X, Xing G, *et al*. The effects of ChatGPT-driven blended teaching model on nursing rounds: A quasi-experimental study. *Nurse Educ Pract* 2025;88:104545.
11 Shin H, De Gagne JC, Kim SS, *et al*. The Impact of Artificial Intelligence-Assisted Learning on Nursing Students' Ethical Decision-making and Clinical Reasoning in Pediatric Care: A Quasi-Experimental Study. *Comput Inform Nurs* 2024;42:704–11.
12 Tseng L-P, Huang L-P, Chen W-R. Exploring artificial intelligence literacy and the use of ChatGPT and copilot in instruction on nursing academic report writing. *Nurse Educ Today* 2025;147:106570.
13 Kejingyun S, Mingjun R. Randomized Controlled Study on the Impact of Problem-Based Learning Combined With Large Language Models on Critical Thinking Skills in Nursing Students. *Nurse Educ* 2025;50:216–20.
14 Lizarondo L, Stern C, Carrier J, *et al*. JBI manual for evidence. Synthesis; 2024.
15 Shamseer L, Moher D, Clarke M, *et al*. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 2015;350:g7647.
16 Rethlefsen ML, Kirtley S, Waffenschmidt S, *et al*. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. *Syst Rev* 2021;10:39.

17 McGowan J, Sampson M, Salzwedel DM, *et al*. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *J Clin Epidemiol* 2016;75:40–6.
18 Kirkpatrick JD, Kirkpatrick WK. Kirkpatrick's Four Levels of Training Evaluation: Association for Talent Development. 2016.
19 Sterne JAC, Savović J, Page MJ, *et al*. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366.
20 Sterne JA, Hernán MA, Reeves BC, *et al*. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355.
21 Hong QN, Fàbregues S, Bartlett G, *et al*. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *EFI* 2018;34:285–91.
22 Guyatt GH, Oxman AD, Schünemann HJ, *et al*. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol* 2011;64:380–2.
23 Lewin S, Booth A, Glenton C, *et al*. Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. *Implement Sci* 2018;13:2.
24 Popay J, Roberts H, Sowden A, *et al*. Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A Product from the ESRC Methods Programme; Lancaster Lancaster University; 2006.
25 Campbell M, McKenzie JE, Sowden A, *et al*. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ* 2020;368.
26 Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res* 2005;15:1277–88.
27 Hoffmann TC, Glasziou PP, Boutron I, *et al*. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 2014;348.
28 Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Med Res Methodol* 2008;8:45.
29 Guetterman TC, Fetters MD, Creswell JW. Integrating Quantitative and Qualitative Results in Health Science Mixed Methods Research Through Joint Displays. *Ann Fam Med* 2015;13:554–61.