Reviewer3

The following are the key revisions that must be addressed before the manuscript can be reconsidered for publication:

1. Revise the PRISMA flow diagram to conform fully to PRISMA 2020 standards, including a complete breakdown of exclusions by reason.

   We updated our literature screening process to strictly follow the PRISMA 2020 guidelines. A fully updated PRISMA flow diagram has been added to the revised manuscript as Figure 1.

2. Include a formal quality appraisal of the included studies using an appropriate tool (e.g., MMAT or CASP checklists) and discuss how study quality influenced the confidence in the synthesized findings.

   We have ensured that the manuscript's methodology and results sections are strictly aligned with the PRISMA 2020 standards, as demonstrated by the newly added PRISMA Flow Diagram. Given that we are addressing a high volume of comments from five independent reviewers, we have prioritized integrating these essential elements directly into the text to ensure clarity and flow.

3. Either justify the inclusion of Studies 3 and 13 (which appear to use non-generative AI systems) with a clear and well-reasoned definition of GenAI, or exclude them and revise the findings accordingly.

   Study 3 and 13 have been excluded.

4. Fix all structural and numbering errors throughout the manuscript. Integrate the Introduction and Related Work sections or clearly differentiate their purpose. Renumber all headings consistently.

   All structural and numbering errors are corrected.

5. Demonstrate explicitly how the SPIDER framework was applied to each study. Consider adding a supplementary analytical matrix showing the coding of all 18 studies across the five SPIDER dimensions.

   We have updated our explanation for SPIDER, which "guided analysis across five dimensions: (1) learner context (e.g., medical students, trainees, clinicians), (2) pedagogical phenomenon of interest (e.g., communication skills, clinical reasoning, engagement), (3) educational design and agent role (e.g., AI as simulated patient, tutor, or instructional avatar), (4) learning mechanisms (e.g., experiential practice, scaffolding, repetition, immersion), and (5) type of evidence reported (e.g., pilot study, mixed-methods, usability evaluation). This framework aims to guide data analysis based on pedagogical perspectives by linking AI agent configurations to learning contexts, educational designs, and underlying learning mechanisms."

6. Substantially revise the Discussion section to deepen the theoretical analysis. Engage more explicitly with established learning theories (e.g., experiential learning, cognitive load, social constructivism) and connect them to specific findings from the reviewed studies.

We have substantially revised the Discussion, especially added relevant discussion on experiential learning, cognitive load, social constructivism. "*The above mechanisms can be further related to established learning theories that many of the reviewed studies have not noted. Firstly, the use of simulated, interactive environments aligns closely with experiential learning theory (Kolb, 1984), where learners experience knowledge being transferred by experience through the "Socratic" conversation with GIVAS. For example, because the "patient" is a virtual AI agent, learners can immediately retry the clinical scenario with a different approach to see how the outcome changes. Secondly, the research findings broadly align with cognitive load theory, as GIVAS may help reduce or add cognitive load depending on the design and functionality of the GIVAS. For example, the use of GIVAS as a tutor or guide to provide learners support to understand clinical documents and generate scenarios can reduce cognitive load. However, research shows that instructional techniques that help novices (e.g., step-by-step guidance) can actually hinder experts, which is seen as the expertise reversal effect (Kalyuga, 2007). Reliance on AI-generated responses may also limit opportunities for deeper cognitive processing and independent reasoning, particularly if learners adopt a passive interaction style. The foundations of social constructivism also reveal that learners may get used to social roles of these virtual agents (i.e., virtual patients) even if they know it is a chatbot, which depends on the extent to which users perceive GIVAS in their interactions as "real social individuals" (Vygotsky & Cole, 1978)."*

7. Revise the scope of conclusions to accurately reflect the preliminary and geographically limited nature of the evidence base. Avoid overstating the strength of findings.

Yes, we have added the scope of conclusions to accurately reflect the preliminary and geographically limited nature of the evidence base in the limitations section. Please see pp.23

> *"In this review, although the findings suggest that GIVAS hold considerable promise in enhancing learning and training in health professions education, the current evidence base remains preliminary. Many of the included studies are small-scale, pilot, or development-focused, with limited longitudinal evaluation. The geographical distribution of studies is relatively narrow (most studies are from developed countries), with evidence concentrated in specific regions and educational contexts. This limits the generalizability of the findings across diverse healthcare systems and learner populations."*

8. Report inter-rater reliability statistics for the screening and data extraction process.

Data extraction was also conducted independently by two reviewers using a standardized data extraction form. Formal inter-rater reliability statistics (e.g., Cohen's kappa) were not calculated, whereas consistency between reviewers was ensured through independent screening and regular discussions to resolve discrepancies. All reviewed studies were critically examined for methodological clarity, relevance to the research questions, and completeness of reported findings during the screening and data extraction process.

9. Engage professional English language editing services and thoroughly revise the entire manuscript for grammar, coherence, and clarity.

We have improved the layout, grammar, and language thoroughly with the help of two native speakers from the authors. We have also undertaken a full proofreading to correct grammatical, stylistic, and spelling errors.

10. Reconsider and expand the Conclusion section to provide substantive synthesis, not just a restatement of the abstract.
We have updated our conclusion. However, due to the word count, we cannot expand too much.