Comparative Efficacy of Generative AI-Based Tutoring vs. Traditional Case-Based Learning in Enhancing Clinical Reasoning: A Systematic Review and Meta-Analysis

Sana Mahdian Rizi[1]*

1. Students Research Committee, Neyshabur University of Medical Sciences, Neyshabur, Iran

*Corresponding author: Mahdiansana@gmail.com



Introduction

Clinical reasoning is a fundamental competency that requires the integration of knowledge with diagnostic synthesis. While Case-Based Learning (CBL) is the gold standard for developing these skills, its implementation is often limited by faculty availability. The emergence of Generative AI (Gen-AI) offers a potential solution for scalable, personalized tutoring. However, empirical evidence regarding its objective impact on reasoning outcomes remains fragmented. This study aimed to synthesize the existing literature to determine whether Gen-AI tutors provide a statistically significant advantage over conventional instructional methods.

Methods

Databases
PubMed | Scopus | Cochrane Library (2020–2025)

Study design
RCTs & quasi-experimental (students)

Outcome
Clinical reasoning proficiency (Script Concordance Tests)

Quality assessment
MERSQI tool (independent reviewers)

Analysis
Random-effects model (RevMan 5.4)

Study Selection

Records identified (n = 1,246) → Identification

↓

After duplicates removed (n = 892) → Screening

↓

Full-text articles assessed (n = 48) → Eligibility

↓

Studies included (n = 14) → Included

Included Studies

14 studies | 1,840 participants | 2020–2025

Results

Overall Effect

SMD = 0.58

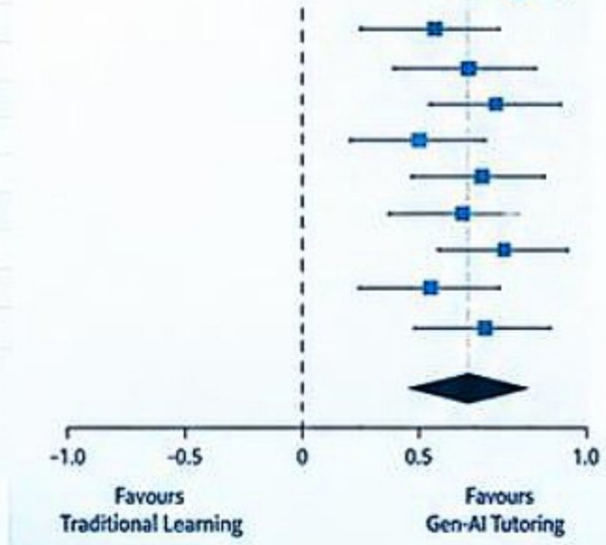
95% CI: 0.34 – 0.82

$p < 0.001$

Gen-AI tutors significantly improved clinical reasoning scores compared to traditional methods.

Forest Plot (Random-Effects Model)

Study	SMD (95% CI)
Study 1	0.72 (0.45 – 0.99)
Study 2	0.38 (0.12 – 0.64)
Study 3	0.64 (0.32 – 0.96)
Study 4	0.51 (0.21 – 0.81)
Study 5	0.78 (0.52 – 1.04)
Study 6	0.46 (0.17 – 0.75)
Study 7	0.59 (0.28 – 0.90)
Study 8	0.67 (0.36 – 0.98)
Study 9	0.42 (0.14 – 0.70)
Study 10	0.61 (0.32 – 0.90)
Study 11	0.55 (0.26 – 0.84)
Study 12	0.70 (0.41 – 0.99)
Study 13	0.47 (0.18 – 0.76)
Study 14	0.63 (0.34 – 0.92)
Total (95% CI)	**0.58** (0.34 – 0.82)



Subgroup Analysis

Interactive / Iterative (Socratic method)	Direct Feedback
SMD = 0.72	SMD = 0.41
95% CI: 0.51 – 0.93	95% CI: 0.18 – 0.64

Heterogeneity

$I^2 = 46\%$ (moderate)

Likely due to variations in LLM architectures and prompt engineering protocols.



Smarter Tutoring. Stronger Reasoning. Better Care.

AI Tutor + Clinical Reasoning + Better Patient Outcomes

Discussion

- Gen-AI tutors may support deeper diagnostic synthesis through deliberate practice and feedback.
- Interactive prompting enhances effectiveness compared to direct feedback.
- Findings support integration of AI as an adjunct, not a replacement, in medical education.

Conclusion

- Gen-AI tutors are effective adjuncts in medical education.
- Further research is needed on long-term retention and ethical frameworks for AI integration in the core curriculum.

Recommendations

- Develop standardized AI tutoring protocols
- Investigate long-term skill retention
- Establish ethical guidelines for AI use in medical education

Keywords: Artificial Intelligence • Clinical Reasoning • Medical Education • Large Language Models • Educational Technology