ORIGINAL ARTICLE

SPOT THE BOT: A comparative quality assessment of AI-generated written instructions for lung ultrasound training

ADRIAN WONG[1,2], NURUL LIANA ROSLAN[3], SERENE HO[4], ROU AN TAN[1], JULINA NOOR[5], GABRIELE VIA[6], FRANCESCO CORRADI[7]

[1]Dept of Intensive Care Medicine, Ng Teng Fong General Hospital, Singapore; [2]Faculty of Medicine, Universiti Malaya, Malaysia; [3]Dept of Emergency Medicine, Hospital Kuala Lumpur, Malaysia; [4]Plain Anaesthesia, United Kingdom; [5]Dept of Emergency Medicine, Faculty of Medicine, Universiti Teknologi MARA, Kuala Lumpur, Malaysia; [6]Cardiac Anesthesia and Intensive Care, Ente Ospedaliero Cantonale (EOC), Istituto Cardiocentro Ticino, Università della Svizzera Italiana (USI), Lugano, Switzerland; [7]Dept of Surgical, Medical, Molecular Pathology and Critical Care Medicine, University of Pisa, Pisa, Italy

ABSTRACT

Background: The rapid proliferation of artificial intelligence (AI) in medical education has outpaced the development of quality assurance methods for AI-generated content. This study provides the first systematic evaluation of AI-generated instructional materials for lung ultrasound (LUS) training.

Methods: The ATLAS study employed a cross-sectional, multi-rater evaluation design comparing six instruction sources (five AI systems and human-generated content) across ten LUS content sessions. Expert evaluators (n=39) assessed materials using five standardized domains: Medical Accuracy, Evidence Completeness, Clarity, Practical Utility, and Pedagogical Quality. Statistical analysis included Kruskal-Wallis tests and pairwise comparisons with Bonferroni correction.

Results: Significant differences existed between instruction sources ($H = 92.582$, $p < 0.001$). Manus AI achieved the highest overall rating (4.55±0.83) and significantly outperformed human instructions in Medical Accuracy ($p = 0.0002$) and Evidence Completeness ($p < 0.001$). Gemini AI (3.94±0.97) performed statistically equivalent to human instructions (4.23±1.00). ChatGPT (2.62±1.35) and Meta (1.53±1.02) performed significantly worse than human instructions ($p < 0.001$). Clarity emerged as the most discriminating criterion with the widest performance range (1.53-4.54).

Conclusions: Certain AI systems can generate high-quality LUS instructional materials that match or exceed human-generated content. However, significant quality variations across AI systems emphasize the critical

Received: 3 January 2025 | Accepted: 23 January 2026

Correspondence: Dr Adrian WONG / Dept of Intensive Care Medicine, Ng Teng Fong General Hospital, Singapore 609606 / E-mail: avkwong@mac.com
ORCID: 0000-0003-4968-7328

importance of systematic evaluation before implementation. These findings support cautious but optimistic integration of high-performing AI systems into medical education with appropriate quality assurance measures.

Key words: artificial intelligence, medical education, lung ultrasound, quality assessment, generative AI, instructional design, point-of-care ultrasound

Introduction

The integration of artificial intelligence (AI) into medical education is rapidly transforming how healthcare professionals are trained, offering unprecedented opportunities for personalized, accessible, and efficient learning (1). From intelligent tutoring systems that adapt to individual learner needs to virtual reality simulations that provide risk-free practice environments, AI is poised to address many of the long-standing challenges in medical pedagogy (2). The recent emergence of sophisticated large language models (LLMs) and other generative AI tools has further accelerated this trend (3), enabling the automated creation of vast amounts of educational content, including detailed instructional materials for complex clinical skills (4). However, this rapid proliferation of AI-generated content has outpaced the development of rigorous methods for quality assurance, raising significant concerns about the accuracy, educational effectiveness, and safety of these novel resources (5, 6).

Lung ultrasound (LUS) has become an indispensable tool in emergency and critical care settings, permitting rapid, non-invasive diagnosis and monitoring of life-threatening conditions such as pneumothorax, pneumonia, pulmonary oedema, and pleural effusion (7). Despite its proven clinical value, widespread adoption of LUS has been hindered by significant challenges in training and competency assessment. A systematic review of the literature by Pietersen et al. revealed a lack of international consensus on LUS education, with most training programs being unstructured and lacking validated assessment methods (8). This heterogeneity in training prevents the consistent, effective application of LUS in clinical practice.

The potential for generative AI to create standardized, evidence-based instructional materials for LUS is immense. AI could, in theory, produce comprehensive curricula that cover everything from basic ultrasound physics to complex clinical integration, tailored to the specific needs of novice learners. However, the unguided use of such tools carries important risks including inaccurate medical information, poor pedagogical structure, or the omission of critical safety considerations, all of which could have direct and detrimental consequences for patient care (9). This underscores the urgent need for a systematic approach to evaluating the quality of AI-generated educational content within the high-stakes environment of medical skills training. To date, no study has systematically evaluated the quality of instructional materials generated by different AI models for a specific, complex clinical skill such as LUS.

The primary objective of this study was to conduct a comprehensive, multi-domain comparative quality assessment of written instructional materials for LUS generated by multiple, distinct AI models. The findings of this study provide the first evidence-based insights into the capabilities and limitations of current generative AI in creating high-quality medical education content, offering crucial guidance for the future development and implementation of AI in clinical training.

Methods

Study design

The ATLAS (Artificial intelligence-generated Training for Lung ultrASound) study employed a cross-sectional, multi-rater evaluation design to assess

the quality of AI-generated written instructional materials for focused LUS training. This study was conducted between June and August 2025 and received ethical approval from the King's College London, UK Research Ethics Committee.

Content generation

AI-generated content

Six distinct sets of written instructional materials for LUS training were generated (supplementary material 1 - 6). Five sets were created by different AI language models; one set was generated by human experts to serve as a control. The AI models used were selected based on their prevalence and distinct underlying architectures at the time of the study:

- **ChatGPT 4.0** (OpenAI, version accessed July 2025)
- **Claude 3 Opus** (Anthropic, version accessed July 2025)
- **Manus 1.5** (Manus, version accessed July 2025)
- **Meta Llama 3** (Meta AI, version accessed July 2025)
- **Gemini 2.5** (Google, version accessed July 2025)

Each AI model was engaged using the same standardized prompt framework (supplement material 7). The prompt was designed to create instructional materials suitable for healthcare professionals with no prior ultrasound experience. It specified the target audience, the ten essential learning sessions for novice LUS training, and the requirement for clear, practical, and evidence-based content. Images or video clips were not included in the training materials.

Human-generated content

The human-generated control training materials were developed by two board-certified intensive care physicians with over 15 years of collective experience in LUS education and curriculum development. The materials were created de novo, referencing the international evidence-based recommendations for point-of-care LUS [7] to ensure they represented a high-quality standard.

Instructional content

Each of the six instruction sets comprised comprehensive educational content covering ten standardized learning sessions essential for novice LUS training:

- Basic Ultrasound Physics
- Probe Selection and Positioning
- Normal LUS Appearance
- Identifying Pleural Sliding
- Recognizing B-Lines
- Detecting Pneumothorax
- Assessing Pleural Effusion
- Lung Consolidation Patterns
- Common Artifacts and Troubleshooting
- Clinical Integration and Decision-Making

Expert evaluator recruitment

Expert evaluators were recruited through professional organizations (WINFOCUS) and institutional networks based on the following criteria:

- **Inclusion Criteria**: Current involvement in LUS education at an institutional or national level; demonstrated expertise through publications, presentations, or recognized training programs; clinical practice involving regular use of LUS.
- **Exclusion Criteria**: Involvement in AI-generated content development for medical education; declared conflicts of interest with AI technology companies; insufficient English proficiency.

Evaluation framework

The evaluation framework was developed following established principles for questionnaire design (10) and competency-based assessment in medical education (11, 12). The use of a structured, multi-domain framework ensures systematic and consistent assessment of instructional material quality (13). The five

domains were chosen to incorporate key principles of instructional design (14) and align with established approaches for assessing ultrasound education (15).

Each content section within every instruction set was evaluated across five standardized domains using a 5-point Likert scale (1 = Poor, 2 = Below Average, 3 = Average, 4 = Good, 5 = Excellent):

1. **Medical Accuracy**: Assessment of factual correctness, adherence to current evidence-based recommendations, and absence of medical errors or misconceptions.
2. **Educational Clarity**: Evaluation of content appropriateness for novice learners, clarity of explanations, logical progression of concepts, and accuracy of sentence structures.
3. **Completeness**: Assessment of comprehensive coverage of essential content, inclusion of critical safety considerations, and adequate depth for competency development.
4. **Practical Utility**: Evaluation of actionable guidance, real-world applicability, hands-on procedural details, and clinical relevance.
5. **Pedagogical Quality**: Assessment of adherence to sound educational principles, effective learning structure, use of appropriate teaching methods, and engagement strategies.

Evaluation protocol

Evaluators completed a structured assessment process for each instruction set, consisting of both quantitative (domain-specific scores from 1 to 5 for each content session) and qualitative (free-text comments and feedback) measures. To ensure assessment independence, evaluators were instructed to complete their grading individually without interaction or discussion with other participants. Evaluators were strictly blinded to the source of the instruction sets (i.e., whether they were generated by an AI model or human experts) and to the identity of other evaluators throughout the process.

Statistical analysis

Statistical analysis was conducted using Prism v 10.4 (GraphPad Software, San Diego, CA). Descriptive statistics were calculated for all variables. Given that the evaluation data were derived from a 5-point Likert scale, which produces ordinal data, non-parametric statistical tests were employed (16). The Shapiro-Wilk test confirmed that the data were not normally distributed.

Overall between-group comparisons of instruction sets were performed using the Kruskal-Wallis test. Pairwise comparisons between the human-generated instructions and each AI system were conducted using Mann-Whitney U tests. Inter-rater reliability was assessed using intraclass correlation coefficients (ICC), with a target ICC > 0.6 considered acceptable (17). Qualitative feedback from free-text comments was analysed using thematic analysis to identify recurring themes. Sample size was determined based on detecting a 1-point difference on the 5-point scale for grading instruction sets, with 80% power and $\alpha = 0.05$. Accounting for multiple comparisons and potential dropout (10%), a minimum of 22 expert evaluators was required. The final sample of 39 evaluators provided a statistical power of >90% for the primary analysis.

Results

Participant characteristics

A total of 39 expert evaluators from four geographic regions completed the ATLAS study assessment. The demographic characteristics of the study participants are presented in Table 1. The majority of respondents were from Malaysia (48.7%, n=19), followed by the United Kingdom (23.1%, n=9). Critical care specialists comprised the largest clinical specialty group (51.3%, n=20), followed by internal medicine physicians (23.1%, n=9). Participants demonstrated moderate experience in LUS training, with a mean experience of 2.49 years (SD = 0.94). 56.4% (n=22) of the participants reported some experience with AI tools, while 41.0% (n=16) reported extensive experience.

Overall quality assessment

A Kruskal-Wallis test revealed a statistically significant difference in the overall quality ratings across the six instruction sources. As summarized in Table 2, Manus AI achieved the highest overall rating (mean = 4.55,

SD = 0.83), followed by human-generated instructions (mean = 4.23, SD = 1.00), Gemini (mean = 3.94, SD = 0.97), Claude (mean = 3.56, SD = 1.11), ChatGPT (mean = 2.62, SD = 1.35), and Meta (mean = 1.53, SD = 1.02). Manus AI also received the highest proportion of high-quality ratings (scores of 4 or 5), with 84.8% of its ratings falling within this category, followed by human instructions (77.1%). In contrast, Meta and ChatGPT received the highest proportion of low-quality ratings (scores of 1 or 2), at 84.4% and 48.7%, respectively.

Domain-specific performance analysis

The performance across the five evaluation domains is presented in Figure 1. Manus AI achieved the highest mean scores across all domains: Medical Accuracy (4.67), Evidence Completeness (4.58), Clarity (4.54), Practical Utility (4.50), and Pedagogical Quality (4.42). Human-generated instructions demonstrated consistent high performance across domains, with scores ranging from 4.18 to 4.42.

Gemini showed moderate performance with relatively consistent scores across domains (range: 3.78-3.99). Claude demonstrated similar moderate performance but with greater variability, particularly a lower score in Clarity (3.40). ChatGPT performed poorly across all domains (range: 2.52-2.87), while Meta achieved the lowest scores in all categories (range: 1.53-2.05).

Table 1. Participant demographics and characteristics.

Characteristic	n (%)
Geographic Distribution	
Malaysia	19 (48.7)
United Kingdom	9 (23.1)
Singapore	6 (15.4)
Other European	5 (12.8)
Clinical Specialty	
Critical Care	20 (51.3)
Internal Medicine	9 (23.1)
Other	10 (25.6)
AI Experience	
Some experience	22 (56.4)
Extensive experience	16 (41.0)

*One respondent did not answer the question on AI experience level.

Comparisons between AI and human-generated instructions

Pairwise comparisons between human-generated instructions and each AI system using Mann-Whitney U tests revealed significant differences for two AI systems after Bonferroni correction ($p < 0.01$), as shown in Table 3. ChatGPT ($U = 203.5$, $p < 0.001$, $r = -0.69$) and Meta ($U = 48.0$, $p < 0.001$, $r = -0.89$) performed significantly worse than human instructions. Notably, Manus ($U = 476.5$, $p = 0.132$, $r = 0.18$) and Gemini ($U = 464.5$, $p = 0.152$, $r = -0.19$) were statistically equivalent to human instructions, indicating these AI systems can generate content of acceptable quality for clinical education purposes.

Pairwise comparisons of overall quality: Human vs AI systems (Table 3)

Domain-specific comparisons revealed that Manus AI significantly outperformed human instructions in Medical Accuracy (mean difference = 0.25, $p = 0.0002$) and Evidence Completeness (mean

Table 2. Overall quality assessment by instruction source.

Source	n	Mean (SD)	Median	High Quality (4-5) n (%)	Low Quality (1-2) n (%)
Manus	33	4.55 (0.83)	5.0	28 (84.8)	1 (3.0)
Human	35	4.23 (1.00)	5.0	27 (77.1)	3 (8.6)
Gemini	33	3.94 (0.97)	4.0	22 (66.7)	1 (3.0)
Claude	32	3.56 (1.11)	4.0	18 (56.3)	6 (18.8)
ChatGPT	39	2.62 (1.35)	3.0	14 (35.9)	19 (48.7)
Meta	32	1.53 (1.02)	1.0	2 (6.3)	27 (84.4)

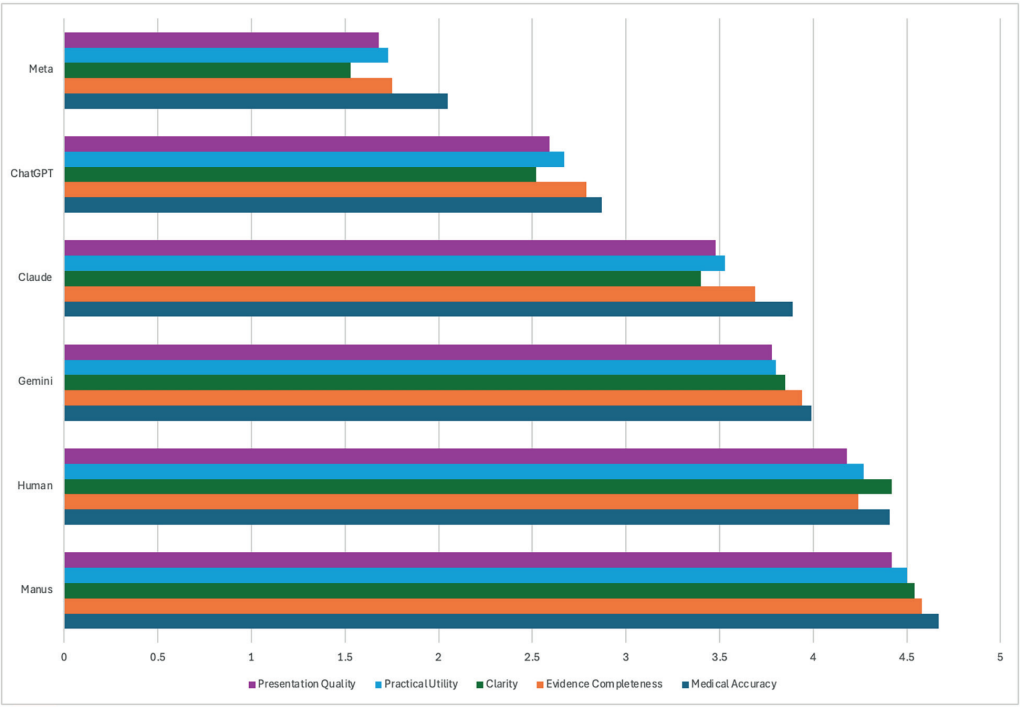

Figure 1. Domain-specific mean scores by instruction source.

Table 3. Pairwise comparisons: Human vs AI systems.

AI System	U-statistic	p-value	Effect Size	Significant Difference
Manus	476.5	0.132	0.184	No
Gemini	464.5	0.152	-0.190	No
Claude	349.0	0.009	-0.368	No*
ChatGPT	203.5	<0.001	-0.693	Yes
Meta	48.0	<0.001	-0.893	Yes

*Not significant after Bonferroni correction.

difference = 0.34, $p < 0.001$). No other significant differences were observed for Clarity, Practical Utility, or Pedagogical Quality.

Reviewer confidence in AI detection (Table 4, Figure 2)

ChatGPT was most confidently identified as AI-generated (71.8% high confidence, mean = 4.03). Conversely, Manus was least confidently identified as AI-generated (57.6% high confidence, mean = 3.58), suggesting it produced more human-like content. Notably, evaluators expressed high confidence (69.4%) that the human-generated instructions were created by AI, indicating a potential bias or difficulty in distinguishing high-quality human writing from sophisticated AI output.

Qualitative analysis

Thematic analysis of the free-text comments revealed several key themes. For high-performing systems like Manus and Gemini, evaluators praised the "clear, concise language," "logical flow," and "comprehensive

evidence base." In contrast, feedback for low-performing systems like ChatGPT and Meta frequently cited issues such as "vague and generic statements," "lack of practical detail," and "medically imprecise language."

Discussion

This study represents the first systematic evaluation of AI-generated instructional materials for focused LUS training, providing critical insights into the current capabilities and limitations of generative AI in medical education. Our findings demonstrate significant heterogeneity in quality across different AI systems, with important implications for the safe and effective implementation of AI-generated content in clinical training. The central finding is that while some AI systems can produce instructional materials that match or even exceed a human-expert standard, others generate content of unacceptably low quality, underscoring the absolute necessity of rigorous, systematic evaluation before deployment.

A key finding of this study is that the Manus AI system not only matched but outperformed human-generated instructions in the domains of Medical Accuracy and Evidence Completeness. This challenges the prevailing assumption that human-generated content is the unequivocal gold standard for medical education. It suggests that an AI system, when appropriately designed and trained on a high-quality, curated knowledge base, can synthesize evidence more comprehensively and with greater fidelity to current guidelines than a human expert working from memory or general knowledge. This has significant implications, suggesting a future role for AI not just as a content creator, but as a tool for evidence synthesis and curriculum validation.

Conversely, the poor performance of ChatGPT and Meta highlights the considerable risks of using general-purpose AI models for specialized medical education without stringent quality control. The qualitative feedback, which noted these systems produced "vague," "generic," and "medically imprecise" content, aligns with broader concerns that some LLMs may prioritize helpfulness and linguistic fluency over factual accuracy (18). While ChatGPT has shown improving performance on standardized knowledge exams (19), our findings suggest that the ability to answer factual questions does not directly translate to the ability to create high-quality, pedagogically sound instructional materials for a complex procedural skill. This distinction is critical for

Table 4. Reviewer confidence in AI detection.

Source	N	Mean (SD)	High Confidence (4-5) (%)
ChatGPT	39	4.03 (0.99)	71.8%
Meta	32	3.97 (0.86)	68.8%
Human	36	3.86 (0.90)	69.4%
Claude	32	3.78 (0.75)	65.6%
Gemini	33	3.76 (0.90)	63.6%
Manus	33	3.58 (0.83)	57.6%

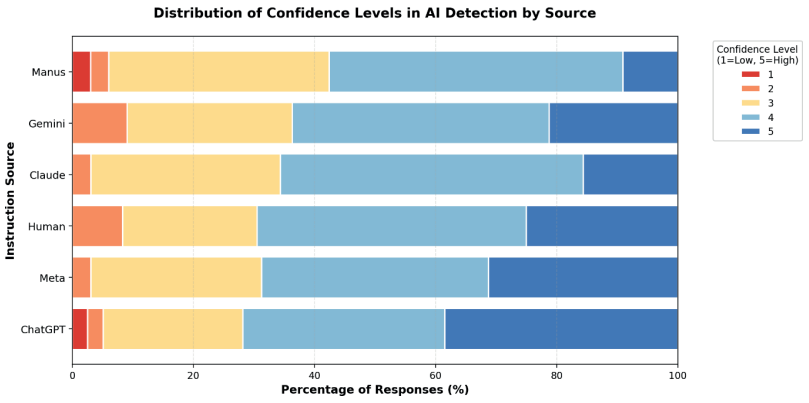

Figure 2. Distribution of confidence level in AI detection by source.

educators to understand. However, a limitation of evaluating proprietary commercial models is their 'black box' nature; the specific algorithms and real-time source selection logic are not publicly disclosed. While Manus appears better optimized for evidence synthesis, the opacity of these systems prevents a definitive technical explanation for the performance disparities observed.

The demonstration that certain AI systems can generate clinically acceptable educational content has implications for global medical education. High-quality LUS training is often limited by the availability of expert instructors and standardized curricula. AI systems that can generate consistent, evidence-based content on demand could democratize access to quality medical education, particularly in low-resource settings or for emerging medical technologies where expert educators are scarce (20, 21).

Our findings align with recent comparative research on AI in medical education while providing novel insights specific to procedural skills training. While other studies have compared the performance of models like Gemini and ChatGPT in answering clinical questions (22, 23), this study is the first to evaluate their output in the context of instructional design. The superior performance of Manus and Gemini in this study suggests that the architectural design and training data of these models may be better suited for generating structured educational content than other general-purpose AI models.

One of the most intriguing findings was the evaluators' difficulty in distinguishing AI from human-generated content. The fact that the human-generated instructions were more frequently misidentified as AI-generated than the Manus and Gemini outputs is a cause for reflection. This may suggest that expert-written educational content can be perceived as formulaic, or that evaluators hold a preconceived bias that high-quality, well-structured content is more likely to be AI-generated. Furthermore, the ability of the highest-performing AI system (Manus) to be the least identified as AI suggests a level of sophistication in its writing style that mimics human nuance- a key factor in creating engaging and effective educational material. This finding has implications for academic integrity, the nature of authorship, and the evolving relationship between human and artificial intelligence in content creation.

Limitations

Several limitations should be considered when interpreting our findings. The study sample, while geographically diverse, was concentrated in certain regions (48.7% from Malaysia) and clinical specialties (51.3% critical care), which may limit generalizability. This concentration of evaluators within specific demographics introduces a potential source of bias, as these groups may share reduced variability in preferred instructional styles or terminology compared to a broader global cohort.

Furthermore, the human-generated control materials were developed by experts from a single specialty (Intensive Care). While these authors possess significant experience, a multidisciplinary panel including respiratory or emergency medicine physicians might have produced control texts with different clinical nuances. Additionally, the use of a standardized prompt structure, while necessary to ensure a fair comparison across models, may have constrained the AI models from demonstrating their full potential in structuring a curriculum de novo. Future research should explore optimal prompt engineering strategies and iterative AI refinement to determine whether different prompt designs yield superior educational outcomes.

Importantly, our study focused on text-based instructions due to the limitations of the current AI models to generate images. Hence, the evaluation was limited to written instructional materials and did not assess multimedia content or interactive elements, which are important components of comprehensive medical education. Additionally, the study assessed content quality but did not evaluate learning outcomes or long-term retention. Finally, the study focused exclusively on LUS, a relatively straightforward point-of-care application; it is possible that different results would occur when more complex skills are the subject matter of the instructional materials studied (24). The rapid evolution of AI models also means these findings represent a snapshot in time, and performance will likely change with future model updates.

Future directions

Future research should focus on longitudinal studies evaluating learning outcomes and skill retention with AI-generated versus human-generated content. Investigation of optimal hybrid approaches combining AI efficiency with human expertise is critical to maximizing educational effectiveness while maintaining quality standards [25]. Additionally, research into domain-specific AI training for medical education could address the performance gaps observed in practical utility and clinical integration. The development of dedicated AI platforms fed with selected subject-specific publications, as suggested by the performance of Manus AI, warrants further investigation.

Conclusion

The ATLAS study provides the first systematic evidence that AI can generate high-quality instructional materials for focused LUS training that meet or, in some cases, exceed the standards of human-generated content. Our comprehensive evaluation of six different instruction sources across five quality domains reveals significant heterogeneity in AI system performance. These findings support a cautious but optimistic integration of high-performing AI systems into medical education, but only when coupled with rigorous, systematic quality assurance measures to ensure the safety and effectiveness of clinical training.

Acknowledgements: NA

Author's contributions: AW came up with the idea and drafted the initial protocol. AW, SH and NR refined the LLM prompts. JN and GV piloted the scoring grid. All authors contributed to the finalising of the submitted manuscript.

Funding: None.

Availability of data and material: Prompt and LLM instructions available as Supplementary Material.

Ethics approval and consent to participate: King's College London REC approval.

Consent for publication: NA.

Competing interests: AW received honorarium for the delivery of educational content from GE, Vygon and Mindray.

Declaration on the use of AI: MS words spelling/grammar checker used in the finalisation of the manuscript.

Supplementary material:

- Supplementary file 1 - ATLAS Set 1.pdf
- Supplementary file 2 - ATLAS Set 2.pdf
- Supplementary file 3 - ATLAS Set 3.pdf
- Supplementary file 4 - ATLAS Set 4.pdf
- Supplementary file 5 - ATLAS Set 5.pdf
- Supplementary file 6 - ATLAS Set 6.pdf
- Supplementary file 7 - ATLAS Prompt.pdf

References

1. Thirunavukarasu AJ, Ting DSJ, Elangova K, Gutierrez L, Tan TF, Tin DSW. Large language model in medicine. Nat Med. 2023 Aug;29(8):1930-1940. https://doi.org/10.1038/s41591-023-02448-8
2. Khakpaki A. Advancements in artificial intelligence transforming medical education: a comprehensive overview. Med Educ Online. 2025 Dec;30(1):2542807. doi: 10.1080/10872981.2025.2542807
3. Zhang K, Meng X, Yan X, Ji J, Liu J, Xu H, et al. Revolutionizing health care: The transformative impact of large language models. J Med Internet Res. 2025 Jan 7;27:e59069. doi: 10.2196/59069.
4. Masters K. Ethical use of Artificial Intelligence in Health Professions Education: AMEE Guide No. 158. Med Teach. 2023 Jun;45(6):574-584. doi: 10.1080/0142159X.2023.2186203.
5. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023 Mar 19;11(6):887. doi: 10.3390/healthcare11060887.
6. Rodger D, Mann SP, Earp B, Savulescu J, Bobier C, Blackshaw BP. Generative AI in healthcare education: How AI literacy gaps could compromise learning and patient safety. Nurse Educ Pract. 2025 Aug;87:104461. doi: 10.1016/j.nepr.2025.104461.

7. Volpicelli G, Elbalbary M, Blaivas M, Lichtenstein DA, Mathis G, Kirkpatrick AW, et al. International evidence-based recommendations for point-of-care lung ultrasound. Intensive Care Med. 2012 Apr;38(4):577-91. https://doi.org/10.1007/s00134-012-2513-4
8. Pietersen PI, Konge L, Laursen CB. Lung ultrasound training: a systematic review of published literature in clinical lung ultrasound training. Crit Ultrasound J. 2018 Sep 3;10(1):23. doi: 10.1186/s13089-018-0103-6.
9. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical information: The case of ChatGPT. Res Sq. 2023 Feb 28:rs.3.rs-2566942.. https://doi.org/10.21203/rs.3.rs-2566942/v1
10. Artino AR, La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE Guide No. 87. Medical Teacher. 2014;36(6):463-474. https://doi.org/10.3109/0142159X.2014.889814
11. Parker E, Totura W, Majewski M, Mukherji J, Tetteh E, Byram S. Developing a roadmap for a competency-based point-of-care ultrasound curriculum. J Educ Perioper Med. 2025 Apr 8;27(1):E741. doi: 10.46374/VolXXVII_Issue1_Parker.
12. Höhne E, Recker F, Dietrich CF, Schäfer VS. Assessment methods in medical ultrasound education: A systematic review. Front Med (Lausanne). 2022 June 9;9:871957. https://doi.org/10.3389/fmed.2022.871957
13. Pearce J, Edwards, D, Fraillon J, Coates H, Canny BJ. Wilkinson, D. . The rationale for and use of assessment frameworks: Improving assessment and reporting quality in medical education. Perspect Med Educ. 2015 Jun;4(3):110-8. https://doi.org/10.1007/s40037-015-0182-z
14. Cheung L. Using an instructional design model to teach medical procedures. Med Sci Educ. 2016;26:175-180. doi: 10.1007/s40670-016-0228-9
15. Damewood SC, Leo M, Bailitz J, Gottlieb M, Liu R, Hoffmann B, Gaspari RJ. Tools for measuring clinical ultrasound competency: A systematic review. AEM Educ Train. 2019 Jul 30;4(Suppl 1):S106-S112. doi: 10.1002/aet2.10368.
16. Sullivan, G. M., & Artino, A. R. (2013). Analysing and interpreting data from Likert-type scales. J Grad Med Educ. 2013 Dec;5(4):541-2. doi: 10.4300/JGME-5-4-18.
17. Faherty, A., Counihan, T., Kropmans, T., Finn, Y. (2020). Inter-rater reliability in clinical assessments: Do examiner pairings influence candidate ratings? BMC Med Educ. 2020 May 11;20(1):147. https://doi.org/10.1186/s12909-020-02009-4
18. Mass General Brigham. Large language models prioritize helpfulness over accuracy in medical contexts. Press Release. 2025. https://www.massgeneralbrigham.org/en/about/newsroom/press-releases/large-language-models-prioritize-helpfulness-over-accuracy-in-medical-contexts
19. Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. Sci Rep. 2024 Apr 23;14(1):9330. https://doi.org/10.1038/s41598-024-58760-x
20. Corrado G, Barral J. Advancing medical AI with Med-Gemini. Google Research Blog. 2024. https://research.google/blog/advancing-medical-ai-with-med-gemini/
21. Halalau A, Holmes B, Rogers-Snyr A, Donisan T, Nielsen E, Cerqueira TL, Guyatt G. Evidence-based medicine curricula and barriers for physicians in training: a scoping review. Int J Med Educ. 2021 May 28;12:101-124. doi: 10.5116/ijme.6097.ccc0.
22. Bahir D, Zur O, Attal L, Nujeidat Z, Knaanie A, Pikkel J, Mimouni M, Plopsky G. Gemini AI vs. ChatGPT: A comprehensive examination alongside ophthalmology residents in medical knowledge. Graefes Arch Clin Exp Ophthalmol. 2025 Feb;263(2):527-536. doi: 10.1007/s00417-024-06625-4. https://doi.org/10.1007/s00417-024-06625-4
23. Salman IM, Ameer OZ, Khanfar MA, Hsieh YH. Artificial intelligence in healthcare education: evaluating the accuracy of ChatGPT, Copilot, and Google Gemini in cardiovascular pharmacology. Front Med (Lausanne). 2025 Feb 19;12:1495378. doi: 10.3389/fmed.2025.1495378.
24. Pangaro L, ten Cate O. Frameworks for learner assessment in medicine: AMEE Guide No. 78. Med Teach. 2013 Jun;35(6):e1197-210.doi:10.3109/0142159X.2013.788789.
25. Kolcu G, Çalişkan SA. Advancing Assessment of Reliability in Clinical Education: A Generalizability Theory Perspective. J Med Educ Curric Dev. 2025 Oct 28;12:23821205251384832. doi: 10.1177/23821205251384832.