Original Paper

Performance of ChatGPT, Claude, and AMBOSS on the European Board of Urology In-Service Assessment and Alignment With the European Association of Urology 2025 Guidelines: Comparative Study

Mohamed Eldaneen[1*], MD; Shaza Hendy[1*], MBBS; Omar Ramadan[2], MD; Panagiotis Nikolinakos[1], MSc, MD; Asif Muneer[3,4], MBChB, MD, PhD, MBA; Hussain M Alnajjar[3,4], BSc, MBBS, MCh; Karl H Pang[3], BSc, MBChB, MSc, PhD

[1]Department of Urology, Chelsea and Westminster Hospital NHS Foundation Trust, London, United Kingdom

[2]East Kent Hospitals University NHS Foundation Trust, Kent, United Kingdom

[3]Division of Surgery and Interventional Science, University College London, London, WC1E 6BT, United Kingdom

[4]University College London Hospitals NHS Foundation Trust, London, England, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Karl H Pang, BSc, MBChB, MSc, PhD
Division of Surgery and Interventional Science
University College London
Gower Street
London, WC1E 6BT, WC1E 6BT
United Kingdom
Phone: 44 20 7679 2000
Email: karlpang@doctors.org.uk

Abstract

Background: Recent advances in AI, particularly large language models, have generated growing interest in their application to medical education and examination preparation. However, the accuracy, reasoning quality, and adherence to clinical guidelines of these tools in postgraduate urology assessments remain unclear.

Objective: This study aimed to evaluate the performance of 3 AI tools, ChatGPT (GPT-4.0), Claude (version 4.5), and AMBOSS, on European Board of Urology (EBU)–style multiple-choice questions, with a particular focus on accuracy, insight, concordance, and adherence to European Association of Urology (EAU) guidelines.

Methods: A total of 200 single-best-answer questions from the EBU In-Service Assessment workbook (2021-2022) were input into each AI model. Models were prompted to select an answer and provide an explanation. Two urologists with post–Fellowship of the Royal College of Surgeons (FRCS) training independently assessed the outputs. Accuracy was defined as correct answer selection. Concordance was defined as the logical alignment between the answer and its explanation. Insight was evaluated across 3 domains—nonobvious deduction, discriminative reasoning, and clinical validity—and was graded as low, moderate, or high.

Results: ChatGPT demonstrated the highest accuracy (171/200, 85.5%), compared to Claude and AMBOSS (both 159/200, 79.5%; P=.14). Concordance was also significantly higher for ChatGPT (190/200, 95%) than for Claude (176/200, 88%) and AMBOSS (152/200, 76%; P<.001). Nonobvious deduction was predominantly low to moderate across all models, reflecting the recall-based nature of many questions. ChatGPT and Claude showed stronger discriminative reasoning, while AMBOSS demonstrated limited exclusion of alternative options. Clinical validity was high overall, with ChatGPT showing the greatest consistency with EAU guidelines. There was substantial agreement between the 2 reviewers (weighted κ coefficient >0.61).

Conclusions: AI tools can achieve high accuracy on EBU-style assessments; however, differences in reasoning quality and guideline adherence are evident. ChatGPT demonstrated superior performance across all evaluated domains, supporting its role as a potential adjunct in postgraduate urology education.

(JMIR Form Res 2026;10:e100148) doi: 10.2196/100148

KEYWORDS

European Board of Urology; EBU; artificial intelligence; AI; ChatGPT; Claude; AMBOSS; European Association of Urology guidelines; EAU guidelines

Introduction

The European Board of Urology (EBU) examination is a high-stakes assessment designed to evaluate core and advanced urological knowledge in trainees approaching completion of specialist training. Success in the EBU examination is often viewed as an indicator of readiness for independent practice and is closely aligned with the knowledge base required for fellowship-level examinations, such as the Fellowship of the Royal College of Surgeons (FRCS) exam [1].

The EBU examination is a 2-part assessment comprising a written theory examination (part 1) and an oral viva examination (part 2). The part 1 written examination consists of 110 single-best-answer multiple-choice questions (MCQs) designed to determine whether candidates meet the minimum knowledge standard defined by the EBU. The examination covers the breadth of urological practice and is structured across core domains, including basic science, oncology, endourology, andrology, functional urology, trauma, and other subspecialty areas. The part 2 examination is an oral viva that assesses clinical reasoning and decision-making through structured case-based discussions [1].

Advances in AI, particularly the development of large language models (LLMs), have generated substantial interest in medical education [2,3]. Despite this promise, the application of LLMs in health care education remains contentious due to concerns regarding factual accuracy, hallucinated outputs, lack of transparency in reasoning, and the risk of overreliance by learners [4]. Although several studies have demonstrated strong LLM performance in general medical and nonmedical professional examinations, including the United States Medical Licensing Examination (USMLE) and legal board assessments [5-7], performance within specialty-specific, higher-order clinical domains, particularly those requiring nuanced decision-making, remains less well characterized [8]. Recent work has suggested that AI systems enhanced with specialty-specific guidelines can achieve high performance on urology board-style questions, highlighting both the potential and limitations of such tools in specialist education [9].

In this study, we aimed to evaluate the performance of 3 widely used AI tools: 2 general-purpose models, ChatGPT (GPT-4.0) and Claude (version 4.5), and 1 medical-specific platform, AMBOSS, on a set of EBU-style urology MCQs. Unlike general LLMs, AMBOSS's content and explanations are curated by medical educators and clinicians, which may enhance factual accuracy and guideline alignment in domain-specific settings [10].

In addition, model outputs were assessed for insight and concordance with expert reasoning, using benchmark comparisons against the European Association of Urology (EAU) guidelines [11].

Methods

This study was performed with reference to the METRICS (Modern Quality Assessment Framework for Evaluating Generative AI Studies in Healthcare) framework [12] (Multimedia Appendix 1).

Study Design

This was a comparative evaluation of 3 AI-based tools, ChatGPT (GPT-4.0; OpenAI), Claude (version 4.5; Anthropic PBC), and AMBOSS (AMBOSS GmbH), on EBU-style MCQs, benchmarked against expert clinician assessment using the 2025 EAU guidelines [8] as a reference.

Data collection was performed between December 8 and 12, 2025. ChatGPT was accessed using a ChatGPT Plus subscription through the web interface [13], Claude through the Claude-4.5 Pro web interface [14], and AMBOSS through a full-access institutional subscription [15].

Questions were entered manually using an identical prompt for all models: "You are sitting the European Board of Urology (EBU) written examination. For the following question, select the single best answer from the options provided. Provide your chosen answer and a brief explanation for your reasoning."

Each question was submitted once in a new, empty conversation to minimize context carryover between responses. No system prompts, custom instructions, or user-modifiable parameters were used. Output variability across repeated runs was not assessed and represents a limitation of the study.

Question Source

The most recent available EBU In-Service Assessment Questions booklet (2021-2022) was requested from the EBU by email, and approval was obtained for its use. This official workbook contains 200 single-best-answer MCQs with predetermined correct answers. The In-Service Assessment and its accompanying workbook are not official past papers of the EBU part 1 written examination but serve as structured preparatory and educational resources reflecting the style and scope of the certification examination and do not operate on a pass or fail basis.

EBU examination outcomes are reportedly based on the mean score and SD of the candidate cohort, although official pass rates are not publicly available. Furthermore, the workbook is an educational resource. Therefore, performance in this study cannot be directly equated with performance on the official EBU part 1 examination.

Assessment

Accuracy was defined as whether the selected answer correctly addressed the question (correct vs incorrect).

Concordance was defined as the logical alignment between the selected answer and its accompanying explanation, consistent with previously described LLM evaluation frameworks.

Concordance was assessed on a per-question basis, with each of the 200 questions scored as either "yes" (in agreement with the expected answer) or "no." The concordance percentage was then calculated.

Insight assessment was adapted from prior studies evaluating the presence of nonobvious, clinically meaningful reasoning in AI-generated responses and was assessed across three domains: (1) nonobvious deduction (reasoning beyond the question stem), (2) discriminative reasoning (active exclusion of alternative options), and (3) clinical validity (factual accuracy and alignment with accepted urological practice and the EAU guidelines). Insight was graded as low (minimal or absent reasoning), moderate (partial reasoning), or high (clear, structured, and multistep reasoning). A 3-tier classification was selected to capture gradations in reasoning quality and explanatory depth that may not be adequately represented by a binary assessment [16,17].

Ratings were independently assigned by 2 assessors (2 post-FRCS urologists, ME and OR), and disagreements were resolved through consensus discussion, with adjudication by a senior reviewer (KHP) where required, consistent with established approaches for expert-rated assessment of complex constructs [18].

Analysis

Descriptive data were presented as numbers and percentages with 95% CI, where appropriate. Paired categorical data were analyzed using the McNemar test for comparisons between 2 related groups and the Cochran *Q* test for comparisons among 3 or more related groups. Statistical significance was defined as $P<.05$. The weighted κ coefficient was used to assess interrater reliability for agreement on concordance and insight categorical data [19]. Results were tabulated, and graphs were plotted using Microsoft Excel (version 16).

Results

The accuracy and concordance performance of the 3 AI models with corresponding 95% CIs are summarized in Table 1. ChatGPT achieved the highest accuracy at 85.5% (171/200, 95% CI 80.6%-90.4%), compared with 79.5% (159/200, 95% CI 73.9%-85.1%) for both Claude and AMBOSS ($P=.14$). All 3 models achieved accuracy scores exceeding 70%, indicating a high level of performance on EBU-style questions.

Table 1. Accuracy and concordance of ChatGPT, Claude, and AMBOSS, with 95% CIs (N=200).

Outcomes and models	Positive responses, n (%; 95% CI)	Cochran Q	*P* value
Accuracy		3.9	.14[a]
ChatGPT	171 (85.5; 80.6-90.4)		
Claude	159 (79.5; 73.9-85.1)		
AMBOSS	159 (79.5; 73.9-85.1)		
Concordance		29.3	*<.001* [a,b]
ChatGPT	190 (95; 92.0-98.0)		*.02* [c]
Claude	176 (88; 83.5-92.5)		*<.001* [b,d]
AMBOSS	152 (76; 70.1-81.9)		*.004* [b,e]

[a]Comparison between the 3 AI tools (Cochran Q test).

[b]Statistically significant ($P<.05$).

[c]Comparison between ChatGPT and Claude (McNemar).

[d]Comparison between ChatGPT and AMBOSS.

[e]Comparison between Claude and AMBOSS.

ChatGPT significantly demonstrated the highest concordance at 95% (190/200, 95% CI 92.0%-98.0%), followed by Claude at 88% (176/200, 95% CI 83.5%-92.5%) and AMBOSS at 76% (152/200, 95% CI 70.1%-81.9%; $P<.001$). There were also statistically significant differences observed between the different AI tools (Table 1). The weighted κ coefficient was 1.0, representing perfect agreement.

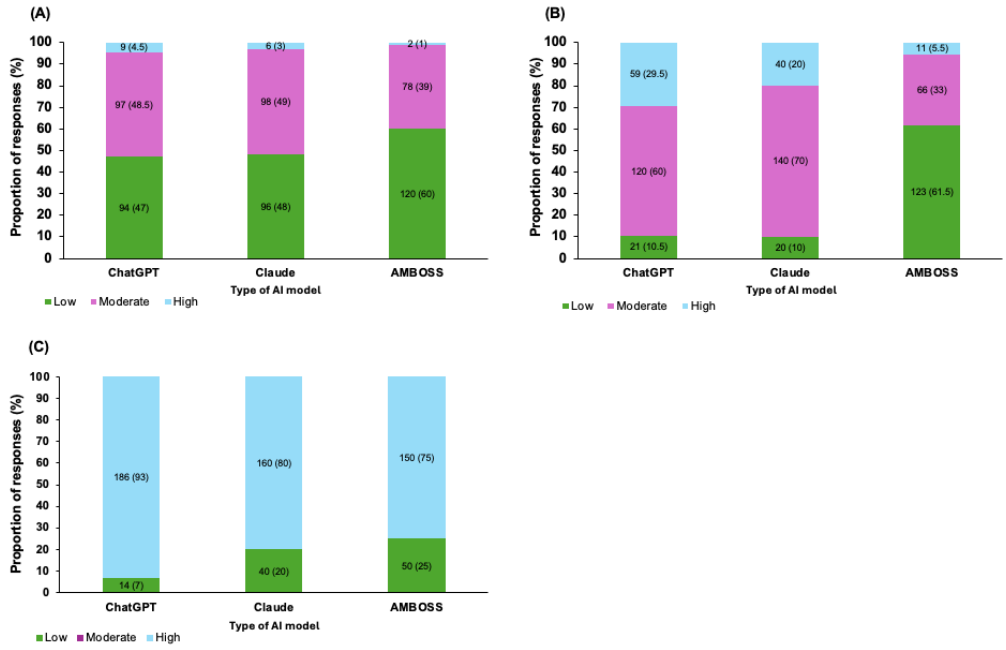
When analyzing subspecialty topics, all 3 models achieved an accuracy of at least 68.1% across all domains (Table 2). An accuracy of 100% was achieved by all models in "miscellaneous" and "transplant and nephrology." ChatGPT and Claude both achieved 100% accuracy in "trauma or emergency." ChatGPT achieved the highest accuracy in "oncology" (82.6%), "pediatric and congenital urology" (90.9%), and "surgical principles" (100%). Claude performed best in "functional urology and benign prostatic hyperplasia" (85.7%) and "lithiasis and infections" (84.6%), while AMBOSS achieved the highest accuracy in "andrology and infertility" (93.8%).

Table 2. Accuracy of ChatGPT, Claude, and AMBOSS across European Board of Urology subspecialty domains.

Domains	ChatGPT, accuracy (%)	Claude, accuracy (%)	AMBOSS, accuracy (%)
Andrology and infertility	81.3	81.3	93.8
Functional urology and benign prostatic hyperplasia	75.0	85.7	75.0
Lithiasis and infections	80.8	84.6	80.8
Miscellaneous	100	100	100
Oncology	82.6	68.1	76.8
Pediatric and congenital urology	90.9	81.8	68.2
Surgical principles	100	86.7	80.0
Transplant and nephrology	100	100	100
Trauma or emergency	100	100	75.0

The results from the insight assessment are demonstrated in Figure 1 and Multimedia Appendix 2. For nonobvious deduction, ChatGPT demonstrated low, moderate, and high ratings in 47% (94/200), 48.5% (97/200), and 4.5% (9/200) of responses, respectively. Claude showed a similar distribution, with low, moderate, and high ratings in 48% (96/200), 49% (98/200), and 3% (6/200) of responses, respectively. AMBOSS demonstrated the lowest performance in this domain, with low, moderate, and high ratings in 60% (120/200), 39% (78/200), and 1% (2/200) of responses, respectively (Figure 1A).

Figure 1. Insight assessment across AI models. (A) Nonobvious deduction, (B) discriminative reasoning, and (C) clinical validity. Bars represent the percentage of 200 questions rated as low, moderate, or high insight for ChatGPT, Claude, and AMBOSS.



For discriminative reasoning, ChatGPT achieved low, moderate, and high ratings in 10.5% (21/200), 60% (120/200), and 29.5% (59/200) of responses, respectively. Claude demonstrated low, moderate, and high ratings in 10% (20/200), 70% (140/200), and 20% (40/200) of responses, respectively. AMBOSS achieved low, moderate, and high ratings in 61.5% (123/200), 33% (66/200), and 5.5% (11/200) of responses, respectively (Figure 1B).

Clinical validity was highest for ChatGPT, with 93% (186/200) of responses rated as high and 7% (14/200) rated as low. Claude achieved high and low ratings in 80% (160/200) and 20% (40/200) of responses, respectively, while AMBOSS achieved high and low ratings in 75% (150/200) and 25% (50/200) of responses, respectively (Figure 1C).

Substantial agreement was observed between the 2 reviewers. The weighted κ coefficient was 0.67, 0.73, and 0.80 for nonobvious deduction, discriminative reasoning, and clinical validity, respectively.

Discussion

Principal Findings

In this study, we evaluated the performance of ChatGPT, Claude, and AMBOSS on 200 EBU-style MCQs using a multidimensional assessment framework that examined not only answer accuracy but also concordance, nonobvious deduction, discriminative reasoning, and clinical validity. Although previous studies have demonstrated that LLMs can achieve pass-level or near–pass-level performance on medical and specialty examinations, these evaluations have largely focused on answer correctness and examination outcomes [20-23]. In contrast, our study assessed the quality and transparency of the reasoning underpinning AI-generated responses, with a particular emphasis on adherence to EAU guideline–based practice.

Previous work supports the growing role of AI in medical education, with GPT-based models demonstrating strong performance across licensing and board examinations [20-23].

However, most studies have focused primarily on examination accuracy rather than the reasoning processes underlying responses. For example, Vaishya et al [24] evaluated AI performance on orthopedic postgraduate examination questions without systematically assessing explanation quality. Similarly, studies of medical question-answering systems and patient-facing chatbots have shown encouraging performance but have also highlighted that answer correctness alone may not adequately capture response quality, transparency, safety, or clinical applicability [25,26].

Assessing guideline adherence is particularly important given the known limitations of AI-generated clinical recommendations. Talyshinskii et al [27] found that although GPT-4.0 demonstrated partially accurate diagnostic knowledge in urolithiasis, its surgical planning recommendations were not consistently aligned with EAU guidelines. Additionally, Eldaneen et al [28] reported that current LLMs provide readily accessible guidance on LUTS; however, their unsupervised use in clinical decision-making remains a concern and may be considered premature.

Broader evaluations of generative AI in health care have similarly highlighted concerns regarding reliability, hallucinations, safety, and the need for human oversight [29,30]. Furthermore, Topol [31] argued that successful integration of AI into health care depends on maintaining human clinical oversight, while systematic reviews have identified ongoing concerns regarding reliability, interpretability, and real-world applicability [32]. Together, these findings emphasize the importance of evaluating not only whether an answer is correct but also whether the underlying reasoning is transparent and clinically appropriate.

Although all 3 models achieved an accuracy exceeding 70%, important differences emerged when reasoning quality was assessed directly. ChatGPT demonstrated the strongest overall performance, with the highest accuracy, concordance, insight scores, and clinical validity. This finding is consistent with recent reports showing strong performance of GPT-4.0–based models across medical knowledge and clinical reasoning assessments [33]. Although all models achieved scores exceeding 70% on the EBU In-Service Assessment workbook, these results should not be interpreted as equivalent to performance on the official EBU part 1 examination because the workbook is an educational resource rather than a validated examination used for certification.

ChatGPT more frequently incorporated structured, multistep reasoning; guideline-based interpretation; and explicit exclusion of alternative answers, resulting in the highest concordance scores.

Claude achieved comparable accuracy and demonstrated strong clinical validity and discriminative reasoning. Its explanations frequently justified why alternative answers were incorrect and were generally consistent with accepted urological practice. However, compared with ChatGPT, Claude was less likely to generate reasoning that extended beyond the information explicitly provided in the question stem.

AMBOSS also demonstrated good overall accuracy and guideline alignment, but its explanations were typically shorter and more descriptive. Although correct answers were often provided, the explanations were less likely to justify the selected option or actively exclude competing alternatives. This was reflected in lower concordance and insight scores than those of ChatGPT and Claude.

High levels of nonobvious deduction were uncommon across all models. This likely reflects the structure of EBU-style questions, many of which assess factual recall and therefore provide limited opportunities for advanced reasoning. Higher ratings were generally observed in questions requiring the integration of multiple clinical concepts or the application of guideline-based management decisions.

The greatest differences between models were observed in discriminative reasoning. ChatGPT and Claude more frequently compared competing answer options and explicitly explained why incorrect alternatives should be excluded. In contrast, AMBOSS often identified the correct answer without fully articulating the reasoning process that led to that conclusion. The ability to actively exclude alternative options represents an important component of both examination performance and clinical decision-making and may enhance the educational value of AI-generated explanations.

All 3 models demonstrated strong clinical validity overall, with most responses aligning with accepted urological practice and EAU guideline recommendations. ChatGPT demonstrated the highest consistency in this domain. These findings are reassuring from an educational perspective, suggesting that clinically inappropriate or potentially misleading recommendations were relatively uncommon across all 3 platforms.

Our findings provide further insight into previously reported limitations of AI performance in medical assessments. Although studies such as those by Sadeq et al [34] have demonstrated variability in chatbot performance across question types and levels of complexity, our results suggest that differences between models are not fully explained by accuracy alone. Models with similar examination performance may differ substantially in

how they justify their answers, exclude alternative options, and maintain alignment with guideline-based practice. This observation is consistent with the work of Holzinger et al [35], who emphasized the importance of evaluating explanations rather than outputs alone.

On the basis of the findings of this study, AMBOSS may be particularly useful for factual revision, whereas ChatGPT and Claude may offer additional value for examination preparation because of their more detailed reasoning processes. Nevertheless, none of the evaluated models should be considered substitutes for clinical supervision or contemporary guideline consultation.

Overall, AI should be introduced in a staged and supervised manner, supporting rather than replacing traditional educational approaches, consistent with recent recommendations regarding the integration of LLMs into medical education [36,37].

Limitations

Several limitations should be acknowledged. This study was limited to single-best-answer MCQs and did not assess performance in more complex or multimodal scenarios, such as imaging interpretation, operative planning, or real-world clinical decision-making. The EBU In-Service Assessment workbook was used as an educational benchmark and is not equivalent to the official EBU part 1 examination; therefore, the findings should not be interpreted as reflecting performance on the certification examination itself. Although predefined criteria, assessor calibration, and dual independent raters were used, some subjectivity is unavoidable. In addition, AI models are dynamic systems whose outputs may change over time because of model updates, and responses were evaluated at a single time point without assessment of repeated-run variability or reproducibility. Potential overlap between model training data and educational resources used in this study cannot be excluded. Furthermore, AI models remain dependent on the scope and currency of their training data, raising concerns regarding alignment with evolving clinical guidelines. Finally, performance in a controlled examination setting may not reflect real-world clinical use.

Conclusions

AI tools demonstrate strong performance on postgraduate urology examination-style questions but differ in reasoning quality and educational value. ChatGPT showed the highest combined accuracy, concordance, and insight, with consistent guideline adherence and strong discriminative reasoning. Claude demonstrated comparable accuracy with good clinical validity but less consistent deductive depth. AMBOSS provided accurate, guideline-aligned responses but relied more on descriptive knowledge with limited explicit reasoning. These findings support the structured and critical use of AI as an adjunct in urology education, emphasizing tools that promote transparent, guideline-based clinical reasoning.

Acknowledgments

The authors declare the use of generative AI (GAI) in the research and writing process. According to the Generative Artificial Intelligence Delegation Taxonomy (2025), the following tasks were delegated to GAI tools under full human supervision: text generation, proofreading and editing, and summarizing text. The GAI tool used was ChatGPT (version 5.5; OpenAI). Responsibility for the final manuscript lies entirely with the authors. GAI tools are not listed as authors and do not bear responsibility for the final outcomes.

Data Availability

All data generated or analyzed during this study are included in this published article.

Funding

The authors declare that no financial support was received for this study.

Authors' Contributions

Data analysis: ME, SH, OR
Data collection: ME, SH
Protocol and project development: KHP
Writing—original draft: ME, SH, KHP
Writing—review and editing: ME, SH, OR, PN, AM, HMA, KHP

Conflicts of Interest

None declared.

Multimedia Appendix 1

Compliance with the METRICS (Modern Quality Assessment Framework for Evaluating Generative AI Studies in Healthcare) checklist.
[DOCX File , 17 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Nonobvious deduction, discriminative reasoning, and clinical validity insight levels across AI models.
[DOCX File , 15 KB-Multimedia Appendix 2]

References

1. Fellow of the European Board of Urology. European Board of Urology. URL: https://www.ebu.com/examination/febu/ [accessed 2026-07-07]
2. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med. Jan 01, 2024;99(1):22-27. [FREE Full text] [doi: 10.1097/ACM.0000000000005439] [Medline: 37651677]
3. Li Q, Qin Y. AI in medical education: medical student perception, curriculum recommendations and design suggestions. BMC Med Educ. Nov 09, 2023;23(1):852. [FREE Full text] [doi: 10.1186/s12909-023-04700-8] [Medline: 37946176]
4. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus. Feb 19, 2023;15(2):e35179. [FREE Full text] [doi: 10.7759/cureus.35179] [Medline: 36811129]
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. Feb 9, 2023;2(2):e0000198. [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
6. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. Feb 08, 2023;9:e45312. [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
7. Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 passes the bar exam. Philos Trans A Math Phys Eng Sci. Apr 15, 2024;382(2270):20230254. [FREE Full text] [doi: 10.1098/rsta.2023.0254] [Medline: 38403056]
8. Touma NJ, Caterini J, Liblk K. Is ChatGPT ready for primetime? Performance of artificial intelligence on a simulated Canadian urology board exam. Can Urol Assoc J. Oct 2024;18(10):329-332. [FREE Full text] [doi: 10.5489/cuaj.8800] [Medline: 38896484]
9. Hetz MJ, Carl N, Haggenmüller S, Wies C, Kather JN, Michel MS, et al. Superhuman performance on urology board questions using an explainable language model enhanced with European Association of Urology guidelines. ESMO Real World Data Digit Oncol. Oct 04, 2024;6:100078. [FREE Full text] [doi: 10.1016/j.esmorw.2024.100078] [Medline: 41646097]
10. Bientzle M, Hircin E, Kimmerle J, Knipfer C, Smeets R, Gaudin R, et al. Association of online learning behavior and learning outcomes for medical students: large-scale usage data analysis. JMIR Med Educ. Aug 21, 2019;5(2):e13529. [FREE Full text] [doi: 10.2196/13529] [Medline: 31436166]
11. Guidelines. European Association of Urology. URL: https://uroweb.org/guidelines [accessed 2026-07-07]
12. Sallam M, Barakat M, Sallam M. A preliminary checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence-based models in health care education and practice: development study involving a literature review. Interact J Med Res. Feb 15, 2024;13:e54704. [FREE Full text] [doi: 10.2196/54704] [Medline: 38276872]
13. ChatGPT. URL: https://chatgpt.com/ [accessed 2026-07-20]
14. Claude. URL: https://claude.ai/login [accessed 2026-07-20]
15. AMBOSS. URL: https://www.amboss.com/int [accessed 2026-07-20]
16. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. Sep 2003;37(9):830-837. [doi: 10.1046/j.1365-2923.2003.01594.x] [Medline: 14506816]
17. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. Sep 1990;65(9 Suppl):S63-S67. [doi: 10.1097/00001888-199009000-00045] [Medline: 2400509]
18. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. Feb 2006;119(2):166.e7-16.e16. [doi: 10.1016/j.amjmed.2005.10.036] [Medline: 16443422]
19. Chmura Kraemer H, Periyakoil VS, Noda A. Kappa coefficients in medical research. Stat Med. Jul 30, 2002;21(14):2109-2129. [doi: 10.1002/sim.1180] [Medline: 12111890]
20. Wang W, Wang B, Zhu Y, Wang Z, Peng S. Evaluation of large language models in medical examinations: a scoping review protocol. PLoS One. Apr 22, 2026;21(4):e0347539. [FREE Full text] [doi: 10.1371/journal.pone.0347539] [Medline: 42018550]
21. Kollitsch L, Eredics K, Marszalek M, Rauchenwald M, Brookman-May SD, Burger M, et al. How does artificial intelligence master urological board examinations? A comparative analysis of different large language models' accuracy and reliability in the 2022 in-service assessment of the European Board of Urology. World J Urol. Jan 10, 2024;42(1):20. [doi: 10.1007/s00345-023-04749-6] [Medline: 38197996]
22. Şahin MF, Doğan Ç, Topkaç EC, Şeramet S, Tuncer FB, Yazıcı CM. Which current chatbot is more competent in urological theoretical knowledge? A comparative analysis by the European Board of Urology in-service assessment. World J Urol. Feb 11, 2025;43(1):116. [doi: 10.1007/s00345-025-05499-3] [Medline: 39932577]

23. 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. Apr 23, 2024;14(1):9330. [FREE Full text] [doi: 10.1038/s41598-024-58760-x] [Medline: 38654011]
24. Vaishya R, Iyengar KP, Patralekh MK, Botchu R, Shirodkar K, Jain VK, et al. Effectiveness of AI-powered chatbots in responding to orthopaedic postgraduate exam questions-an observational study. Int Orthop. Aug 2024;48(8):1963-1969. [doi: 10.1007/s00264-024-06182-9] [Medline: 38619565]
25. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med. Mar 2025;31(3):943-950. [doi: 10.1038/s41591-024-03423-7] [Medline: 39779926]
26. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. Jun 01, 2023;183(6):589-596. [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
27. Talyshinskii A, Juliebø-Jones P, Zeeshan Hameed BM, Naik N, Adhikari K, Zhanbyrbekuly U, et al. ChatGPT as a clinical decision maker for urolithiasis: compliance with the current European Association of Urology guidelines. Eur Urol Open Sci. Sep 16, 2024;69:51-62. [FREE Full text] [doi: 10.1016/j.euros.2024.08.015] [Medline: 39318971]
28. Eldaneen M, Eissa A, Sabaa M, Nikolinakos P, Saber-Khalaf M, Pang KH. Accuracy of large language models in answering urological questions on lower urinary tract symptoms: comparison with the EAU 2025 guidelines. Cent European J Urol. 2026;79(2):152-160. [doi: 10.5173/ceju.2026.0011] [Medline: 42375715]
29. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. Mar 30, 2023;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
30. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. May 4, 2023;6:1169595. [FREE Full text] [doi: 10.3389/frai.2023.1169595] [Medline: 37215063]
31. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. Jan 2019;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
32. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). Mar 19, 2023;11(6):887. [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
33. Bicknell BT, Butler D, Whalen S, Ricks J, Dixon CJ, Clark AB, et al. ChatGPT-4 Omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Med Educ. Nov 06, 2024;10:e63430. [FREE Full text] [doi: 10.2196/63430] [Medline: 39504445]
34. Sadeq MA, Ghorab RM, Ashry MH, Abozaid AM, Banihani HA, Salem M, et al. AI chatbots show promise but limitations on UK medical exam questions: a comparative performance study. Sci Rep. Aug 14, 2024;14(1):18859. [FREE Full text] [doi: 10.1038/s41598-024-68996-2] [Medline: 39143077]
35. Holzinger A, Carrington A, Müller H. Measuring the quality of explanations: the System Causability Scale (SCS): comparing human and machine explanations. Kunstliche Intell (Oldenbourg). 2020;34(2):193-198. [FREE Full text] [doi: 10.1007/s13218-020-00636-z] [Medline: 32549653]
36. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. Jun 01, 2023;9:e48291. [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
37. Masters K. Artificial intelligence in medical education. Med Teach. Sep 2019;41(9):976-980. [doi: 10.1080/0142159X.2019.1595557] [Medline: 31007106]

Abbreviations

EAU: European Association of Urology
EBU: European Board of Urology
FRCS: Fellowship of the Royal College of Surgeons
LLM: large language model
MCQ: multiple-choice question
METRICS: Modern Quality Assessment Framework for Evaluating Generative AI Studies in Healthcare
USMLE: United States Medical Licensing Examination

Edited by J Sarvestan; submitted 03.May.2026; peer-reviewed by P-Y Cheng; comments to author 01.Jun.2026; revised version received 02.Jul.2026; accepted 03.Jul.2026; published 02.Sep.2026

Please cite as:
Eldaneen M, Hendy S, Ramadan O, Nikolinakos P, Muneer A, Alnajjar HM, Pang KH
Performance of ChatGPT, Claude, and AMBOSS on the European Board of Urology In-Service Assessment and Alignment With the European Association of Urology 2025 Guidelines: Comparative Study
JMIR Form Res 2026;10:e100148
URL: https://formative.jmir.org/2026/1/e100148
doi: 10.2196/100148
PMID: