Urologia Internationalis

Research Article

Urol Int 2024;108:359–366
DOI: 10.1159/000537854

Received: November 20, 2023
Accepted: January 17, 2024
Published online: March 30, 2024

Evaluating the Efficacy of AI Chatbots as Tutors in Urology: A Comparative Analysis of Responses to the 2022 In-Service Assessment of the European Board of Urology

Matthias May[a] Katharina Körner-Riffard[b] Lisa Kollitsch[c] Maximilian Burger[b] Sabine D. Brookman-May[d, e] Michael Rauchenwald[c, f] Martin Marszalek[c] Klaus Eredics[c, g]

[a]Department of Urology, St. Elisabeth Hospital Straubing, Brothers of Mercy Hospital, Straubing, Germany; [b]Department of Urology, Caritas St. Josef Medical Centre, University of Regensburg, Regensburg, Germany; [c]Department of Urology and Andrology, Klinik Donaustadt, Vienna, Austria; [d]Department of Urology, University of Munich, LMU, Munich, Germany; [e]Johnson and Johnson Innovative Medicine, Research and Development, Spring House, PA, USA; [f]European Board of Urology, Arnhem, The Netherlands; [g]Department of Urology, Paracelsus Medical University, Salzburg, Austria

Keywords
Artificial intelligence · Large language model · Urology · In-Service Assessment · European Board of Urology · ChatGPT · Bing AI

Abstract

Introduction: This study assessed the potential of large language models (LLMs) as educational tools by evaluating their accuracy in answering questions across urological subtopics. ***Methods:*** Three LLMs (ChatGPT-3.5, ChatGPT-4, and Bing AI) were examined in two testing rounds, separated by 48 h, using 100 Multiple-Choice Questions (MCQs) from the 2022 European Board of Urology (EBU) In-Service Assessment (ISA), covering five different subtopics. The correct answer was defined as "formal accuracy" (FA) representing the designated single best answer (SBA) among four options. Alternative answers selected from LLMs, which may not necessarily be the SBA but are still deemed correct, were labeled as "extended accuracy" (EA). Their capacity to enhance the overall accuracy rate when combined with FA was examined. ***Results:*** In two rounds of testing, the FA scores were achieved as follows: ChatGPT-3.5: 58% and 62%, ChatGPT-4: 63% and 77%, and BING AI: 81% and 73%. The incorporation of EA did not yield a significant enhancement in overall performance. The achieved gains for ChatGPT-3.5, ChatGPT-4, and BING AI were as a result 7% and 5%, 5% and 2%, and 3% and 1%, respectively ($p > 0.3$). Within urological subtopics, LLMs showcased best performance in Pediatrics/Congenital and comparatively less effectiveness in Functional/BPS/Incontinence. ***Conclusion:*** LLMs exhibit suboptimal urology knowledge and unsatisfactory proficiency for educational purposes. The overall accuracy did not significantly improve when combining EA to FA. The error rates remained high ranging from 16 to 35%. Proficiency levels vary substantially across subtopics. Further development of medicine-specific LLMs is required before integration into urological training programs.

Matthias May and Katharina Körner-Riffard shared first authorship.

karger@karger.com
www.karger.com/uin

Correspondence to:
Katharina Körner-Riffard, k.koerner@gmx.de

Introduction

Artificial Intelligence (AI) refers to the use of algorithms to simulate human cognition and reasoning [1]. Large language models (LLMs) are a subset of AI that have been trained to understand and generate human language and thus have emerged as valuable tools to provide tailored content quickly on a wide range of topics [1, 2].

Within the ever-evolving landscape of educational innovation, a shift from traditional textbooks to LLMs for knowledge acquisition is on the horizon. This transition underscores the potential value of LLMs as valuable tools for medical information and education, given to their interactive capabilities.

Nevertheless, an in-depth exploration is warranted to determine the feasibility of effectively integrating AI Chatbots into medical education. Several studies have indicated that LLMs can achieve the passing threshold in various medical exams [3–8]. However, current evidence revealed significant variability in accuracy among different LLMs after repeated testing, highlighting a concerning lack of reliability [4].

Scarce data are available regarding the knowledge of various LLMs in the field of urology [3, 4, 9–11]. Currently, only one study distinguished between general urological and uro-oncological questions in urology exams, indicating a paucity of data on the performance of LLMs in specific urological subfields [3]. In this analysis, we evaluated the performance of ChatGPT-3.5, ChatGPT-4, and Bing AI across five urological subtopics using 100 multiple-choice questions (MCQs) from the 2022. In-Service Assessment of the European Board of Urology.

The European Board of Urology (EBU) In-Service Assessment (ISA) traditionally considers only one single best answer out of four options as correct, referred to as "formal accuracy" (FA). In addition to investigating the FA of LLMs, our study aimed to explore the possibility of considering responses by LLMs initially marked as incorrect as academically appropriate, a concept referred to as "extended accuracy" (EA). Prior research has predominantly focused on distinguishing between correct and incorrect answers, often overlooking nuanced alternatives. Our study goes beyond FA, evaluating the informational content and accuracy of formally incorrect responses (EA) and assessing whether the resulting achieved gains contribute to an enhanced overall rate of correct answers.

Materials and Methods

Utilized LLMs

1. ChatGPT-3.5 (accessed via: https://chat.openai.com/):
   ChatGPT is a state-of the-art AI language model created by OpenAI, released in November 2022. It has been trained on extensive text data, allowing to comprehend human-like text, answer questions, and engage in text-based conversations across various topics. ChatGPT's knowledge is limited to information available up to its training cutoff date in September 2021.
2. ChatGPT-4 (accessed via: https://chat.openai.com/):
   ChatGPT-4, launched in March 2023, is an evolutionary upgrade from its predecessor, with improved abilities in context understanding and generating coherent responses. It retrieves up-to-date information from Internet search engines, making it promising for domain-specific inquiries.
3. Bing AI (accessed via: https://www.bing.com/?/ai):
   Bing AI is an AI-powered search tool developed by Microsoft and OpenAI, introduced in February 2023. It leverages the Microsoft Prometheus model, built upon OpenAI's GPT-4 LLM. The Bing core search ranking engine is constantly improved by this AI model. It utilizes a transformer-based semantic ranking engine to grasp the underlying meaning of text. Notably, Bing AI has the capability to cite its information sources, and the information it provides is current as of the time of the search.

Question Administration and Study Endpoints

The study was conducted with the approval and cooperation of the EBU. The ISA-2022 questions were exclusively and confidentially provided by the EBU for this study, thereby eliminating any possibility of prior training of LLMs.

In August 2023, we sequentially inputted all 100 MCQs from ISA-2022 into the LLMs in the following order: ChatGPT-3.5, ChatGPT-4, and Bing AI. Two rounds of questioning were conducted for each LLM, with a minimum 48-h gap in between. This process included presenting each question with the prompt: "Hey LLM, please answer the following single correct answer multiple choice question (correct answers are A, B, C, or D)": followed by the question itself and the four answer options labeled A to D.

All responses were recorded in a binary form (incorrect vs. correct) following the answers classified as correct by the EBU (FA). Two experienced board-certified urologists (K.E. and K.K.-R.) administered the respective queries. Besides recording the FA outcome, urological specialists made a consensus decision in each round for each LLM to assess whether the answers, initially labeled as incorrect, may possess conceptual and academic value (EA). The study's primary objective was to quantify the increase in correct responses per round and per LLM and determine its statistical significance. The secondary objective was to evaluate the performance of the three LLMs across various urological subtopics, encompassing:

Oncology (n = 35), Functional Urology, Benign Prostate Syndrome (BPS) and Incontinence (n = 14), Urolithiasis and Infection (n = 15), Pediatrics and Congenital (n = 11), and mixed themes (n = 25). In every round, FA and EA were compared for each LLM and subtopic. The answer categories directly corresponded to subtopics in the EBU-ISA exams, which use this categorization. Since this study collaborated with the EBU, it was consistent to evaluate the LLMs according to the clinically relevant categories defined by the board to allow direct comparison to areas of expertise assessed in urological training and certification.

Statistical Analysis

Results were summarized using frequencies and proportions for categorical variables, compared through χ^2 test or Fisher's exact test. Significance was set at $p \leq 0.05$ for all two-tailed tests. For post

Table 1. Improvements in FA and EA by three LLMs in two rounds

	Rd. 1 - FA, *n* (%)	Rd. 1 - EA, *n* (%)	*p* value	Rd. 2 - FA, *n* (%)	Rd. 2 - EA, *n* (%)	*p* value
ChatGPT-3.5	58 (58)	65 (65)	0.309	62 (62)	67 (67)	0.460
ChatGPT-4	63 (63)	68 (68)	0.457	77 (77)	79 (79)	0.733
BING AI	81 (81)	84 (84)	0.577	73 (73)	74 (74)	0.873

EA, extended accuracy; FA, formally accuracy; Rd., round.

hoc analyses related to the second study objective, we adjusted the alpha level using Bonferroni correction. This involved conducting five tests for each comparison, leading to a revised threshold for statistical significance set at $p \leq 0.01$. Statistical analysis was performed using SPSS 29.0 (IBM Corp., Armonk, NY, USA).

Results

Table 1 presents the FA and EA separately for each LMM in each round. In the two rounds, the FAs for ChatGPT-3.5 were 58% and 62%, for ChatGPT-4 63% and 77%, and for BING AI 81% and 73%. When incorporating EA alongside FA, ChatGPT-3.5 improved by 7% and 5% in the two rounds (Fig. 1a), Chat-GPT-4 achieved gains of 5% and 2% (Fig. 1b), and Bing AI enhanced its correct response rate by 3% and 1% (Fig. 1c). All achieved gains were statistically insignificant (always $p > 0.3$, Table 1).

Table 2 displays the FA and the EA for each round and each LLM, subdivided by urological subtopics. Significant differences were found for the FA of Chat-GPT-3.5 in the second round ($p = 0.030$) and the FA of Bing AI in the second round ($p = 0.049$), as well as for the EA of ChatGPT-4 in the first ($p = 0.026$) and second rounds ($p = 0.048$). In all other rounds, differences in FA and EA among subtopics were not statistically significant for all LLMs ($p > 0.05$). Post hoc analyses revealed no significant differences in EA across subtopics for ChatGPT-3.5. However, a higher FA was observed for the Pediatrics/Congenital subtopic in the second round, though it did not reach the corrected significance level ($p = 0.035$).

In ChatGPT-4's first round, Pediatrics/Congenital questions showed an increase in FA, but did not reach the corrected significance level ($p = 0.045$). In the same round, a higher EA tendency was found in Urolithiasis/Infection ($p = 0.021$), while the second round indicated lower EA for Functional/BPS/Incontinence ($p = 0.027$).

Bing AI displayed a significant decrease in FA during the second round in the subtopic Functional/BPS/Incontinence ($p = 0.006$). Furthermore, there was an observable trend toward reduced EA in the second round for Functional/BPS/Incontinence ($p = 0.027$), although it did not reach the corrected significance level.

Discussion

This study surpasses the scope of mere "FA" and delves into the assessment of informational content and accuracy of apparently incorrect responses. The concept of error rates does not invariably signify a proportion of incorrect responses; rather it indicates that the "optimal" answer was not consistently identified. Furthermore, it doesn't necessarily indicate that alternative answers are categorically incorrect. This perspective, previously unexplored in prior studies, is first defined in this present research. Consequently, it offers crucial insights into current limitations of LLMs in acquiring urological knowledge. The absence of significant improvements between rounds contradicts the expectation that repetitive use would enhance the applicability of medical education. With ISA error rates ranging from 16 to 35%, LLMs currently fall short of achieving satisfactory urological understanding necessary for effective training. Notably, evaluating EA did not result in significantly enhanced accuracy. Nevertheless, we recommend that future research on LLMs performance considers overall response accuracy (combining FA and EA) rather than solely focusing on FA, to assess LLMs' performance. This comprehensive approach provides a more precise representation of error rates and yields valuable insights in any medical field. It is imperative to ascertain the frequency at which these AI Chatbots provide inaccurate information, as this serves as the genuine benchmark for their utilization in educational settings. Our findings align with other studies that demonstrated knowledge gaps in LLM within the field of medicine [3–7, 9] and variability across subtopics within a specialty [3, 7, 12–15].

Comparable studies have indicated that ChatGPT-3 exhibits subpar performance in responding to queries

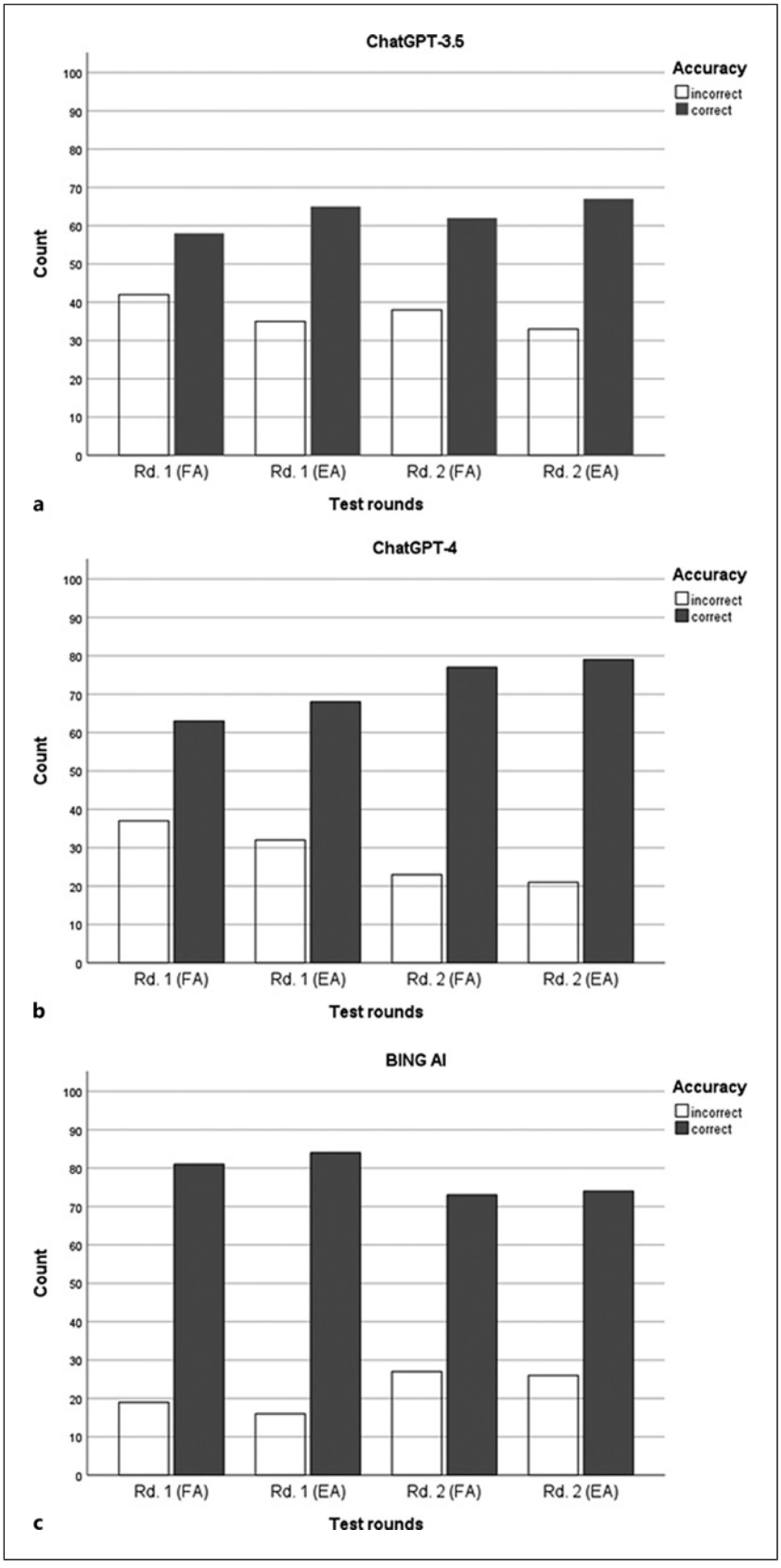

Fig. 1. Representation of the increase in correct responses per round from each LLM across two question rounds. The responses are evaluated based on their FA, and the EA provided by the respective LLM. Panels **a–c** illustrate the two question rounds of the LLMs ChatGPT-3.5, ChatGPT-4, and BING AI, respectively. EA, extended accuracy; FA, formally accuracy; Rd., round.

Table 2. Percentage of correct responses by each LLM, segmented by urological subtopics

LLM	Oncology (n = 35), %	Functional/ BPS/ incontinence (n = 14)	Urolithiasis/ infection (n = 15)	Pediatrics/ congenital (n = 11)	Mixed themes (n = 25)	p value
ChatGPT-3.5 (Rd. 1 - FA)	57.1	42.9	73.3	81.8	48	**0.174**
ChatGPT-3.5 (Rd. 1 - EA)	62.9	64.3	80	81.8	52	**0.310**
ChatGPT-3.5 (Rd. 2 - FA)	51.4	50	86.7	90.9	56	**0.030**
ChatGPT-3.5 (Rd. 2 - EA)	57.1	64.3	86.7	90.9	60	**0.107**
ChatGPT-4 (Rd. 1 - FA)	54.3	42.9	80	90.9	64	**0.058**
ChatGPT-4 (Rd. 1 - EA)	57.1	50	93.3	90.9	68	**0.026**
ChatGPT-4 (Rd. 2 - FA)	80	57.1	86.7	100	68	**0.074**
ChatGPT-4 (Rd. 2 - EA)	80	57.1	93.3	100	72	**0.048**
Bing AI (Rd. 1 - FA)	71.4	78.6	73.3	100	92	**0.123**
Bing AI (Rd. 1 - EA)	77.1	85.7	73.3	100	92	**0.214**
Bing AI (Rd. 2 - FA)	80	42.9	66.7	90.9	76	**0.049**
Bing AI (Rd. 2 - EA)	80	50	66.7	90.9	76	**0.138**

BPS, benign prostate syndrome; EA, extended accuracy; FA, formally accuracy; Rd., round. The assessment of LLM occurred in two rounds, with each round classifying the responses as either formally accurate or extendedly accurate.

within the American Urological Association (AUA) Self-assessment Study Program (SASP) [10, 11]. Deebel et al. [10] administered a total of 268 questions to ChatGPT-3. It exhibited improved performance on the 2021 AUA-SASP questions (42.3% correct) compared to the 2022 set (30% correct). Findings of Huynh et al. [11] revealed that ChatGPT-3 only answered 38 out of 135 (28.2%) MCQ and 36 out of 135 (26.7%) open-ended questions accurately. Response regeneration reduced the number of indeterminate answers, but successive regenerations did not significantly improve the accuracy of responses, as the majority remained incorrect, for both open-ended questions and MCQs [11]. The LLM "Uro_chat," based on ChatGPT-3.5-Turbo and specifically trained on the European Urological Guidelines, delivered accurate responses to nearly two-thirds of the 2022 ISA of the EBU [3]. However, in a comparative analysis of ChatGPT-3.5 and ChatGPT-4 performance in the 2022 EBU-ISA, both models demonstrated lower accuracy rates compared to human examinees [8].

Correct response rates were inversely related to question difficulty, with ChatGPT-4 performing better on more challenging questions. The reliability of LLMs in assessing question difficulty and their performance showed a significant variation across different testing rounds [4, 8]. To investigate improving LLM knowledge bases, a follow-up testing round was conducted 10 weeks after an initial comparative study of ChatGPT-3.5, ChatGPT-4, and Bing AI. This supplementary evaluation reassessed their performances using the 2022 ISA-EBU

Questions. However, the results did not reveal any significant performance enhancements across the tested LLMs [4]. This underscores the current limitations in the adaptive learning abilities of LLMs, as they are confined to responding based on the data available at the time of their last training. It highlights the importance of regularly updating, continuously training, and meticulously maintaining LLMs, particularly for their application in the acquisition of medical knowledge.

In the current analysis, the tested LLMs achieved acceptable accuracy rates, but post hoc subgroup analysis revealed notable variations and gaps in their knowledge levels. LLMs excelled in pediatric-urology questions but showed lower proficiency in functional urology and moderate success in oncology, suggesting limitations in applicability across urological subtopics. The LLMs' superior performance in pediatrics/congenital medicine over functional medicine/BPS/incontinence could be attributed to training data bias, as it may have been trained more extensively on pediatric topics. Furthermore, greater accessibility to clinical data in pediatrics/congenital urology likely facilitated the LLMs' learning, compared to limited data in functional medicine/BPS/incontinence. The intricate nature of concepts within functional medicine/BPS/incontinence may present challenges for LLMs, particularly when lacking sufficient training encompassing these nuanced aspects.

In the future, it will be essential to develop specially trained LLMs tailored to specific subject areas for effective knowledge acquisition. Further studies should focus on subgroup analyses within medical specializations, addressing the challenge of inconsistent responses due to the lack of reliability in LLMs [4]. Repetitive assessments are crucial to ensure the accuracy of LLM-generated responses. In contrast, studies have evaluated LLMs using frequently asked questions from patients and recommendations from the European Association of Urology (EAU) Guidelines. Caglar et al. [9] found that ChatGPT provided satisfactory responses in pediatric urology inquiries, particularly in the fields of phimosis, hypospadias, acute scrotum, and vesicoureteral reflux. Additionally, another study demonstrated that ChatGPT-3 correctly answered 94.6% of questions related to urolithiasis [15]. In both studies, open-ended questions were used as input [9, 15]. These questions were likely to be generally formulated and may not reach the complexity and clinical thinking required in EBU-ISA questions, as they were posed by laypeople.

In a cross-sectional analysis by Musheyev et al. [16], accuracy, understandability, and actionability of responses generated by four different LLMs were evaluated, concerning the top five search queries related to prostate, bladder, kidney, and testicular cancers according to Google trends. Their results indicate that while AI Chatbots may provide accurate and high-quality responses, the use of complex medical terminology could be challenging for users without medical background. Notably, the informational quality was found to be 80%, with no reports of misinformation [16].

When compared to other social media platforms such as TikTok, YouTube, and Instagram, LLMs appear to be a more reliable source of medical information [17]. These findings support the potential future integration of AI Chatbots in patient education. However, ongoing monitoring and evaluating of these tools are essential to prevent misinformation, which is evidently present in the search databases LLMs are applying.

Recently, two studies have shifted focus, investigating ChatGPT's ability to generate qualitative MCQs for medical examinations [18, 19]. Klang et al. [18] employed ChatGPT-4 to write a 210 MCQs examination based on an existing exam template. After revision by specialist physicians who were blinded to the source of the questions, only one MCQ generated by ChatGPT-4 was defined as false and 15% of questions required corrections. The authors concluded that GPT-4 can serve as a supplementary tool in generating MCQs for medical examinations; however, thorough scrutiny by specialist physicians remains essential.

Cheung et al. [19] compared 50 MCQs generated by ChatGPT to 50 MCQs drafted by two university professors with reference to two standard undergraduate medical textbooks. It took ChatGPT only 20 min, whereas the two professionals required 211 min to create the 50 questions for a medical graduate exam. Blinded assessors found satisfactory performance in AI-generated questions across different domains. However, AI scored lower in relevance compared to human-created questions. No significant differences were observed in question quality in other areas or in total scores.

In the future, AI has the potential to greatly save time by offering personalized learning experiences adapted to individual styles and paces. Moreover, it assists in navigating extensive medical literature, facilitating knowledge acquisition, and expediting the creation of study materials or educational presentations. This increased efficiency enables medical professionals to concentrate more on comprehending and applying knowledge, rather than dedicating extensive time to information retrieval.

An investigation into the current landscape of LLM utilization among urologists delves into the usage, opinions, and experiences of practitioners worldwide

regarding ChatGPT [20]. In a web-based survey, 47.7% of the 456 participating urologists reported using ChatGPT or other LLMs in their academic practice. Additionally, 19.8% mentioned incorporating AI in their clinical practice. Notably, 62.2% expressed concerns about potential ethical issues when utilizing ChatGPT for scientific or academic writing, and 53% encountered limitations in the application of ChatGPT within their academic practices. Commonly reported limitations included inaccuracies (44.7%), lack of specificity (42.4%), and response variability (26.5%) [20].

The World Health Organization (WHO) outlined an ethical framework for using LLMs in healthcare and medicine, defining that maintaining human control over healthcare decisions and providers' authority using valid patient information is crucial. Protecting data privacy and requiring informed consent is also essential. Public consultation and debate providing sufficient information should precede any AI system design or deployment [21]. If these ethical standards can be upheld and if LLMs can be technically enhanced or trained to consistently demonstrate up-to-date and guideline-compliant medical knowledge, AI will play a crucial role in medicine and urology. Modern technology has the potential not only to support education but also to save time in clinical settings by managing bureaucratic tasks. As LLMs excel in rapidly processing large amounts of data, making them valuable for summarizing and organizing patient information is based on specific queries.

AI holds tremendous potential in fields like education and healthcare, but we are still in the early stages of this technological revolution. Developing AI, especially in sensitive areas like medicine, demands careful planning, rigorous testing, and continuous refinement. Addressing ethical considerations, data privacy, and system reliability is crucial. The journey ahead is exciting, with each step bringing us closer to realizing the full potential of AI.

This study has several limitations. First, the results should be considered within a temporal context since LLMs are continuously evolving and expected to improve with better databases. Second, only three types of LLM were tested, and novel or medically trained LLMs were not included. ChatGPT 3.5, ChatGPT 4, and Bing AI were selected as they represented the most widely used language models (LLMs) during the conception of this study. ChatGPT was the pioneering and widely adopted LLM in this domain. Bing AI was chosen due to Microsoft's extensive reach. Other models, such as Google's PALM-Med2, necessitate substantial resources and were deemed unfeasible within the scope of this study. Additionally, the exclusive urology-specific LLM, Uro_Chat, is currently temporarily unavailable. Therefore, these three commonly accessible LLMs were a logical choice for initially exploring the capabilities of medical question answering in this research.

Third, the utilized questions were limited to MCQ format with a single best answer design, and no other question designs were evaluated. Additionally, the 100 MCQs may not fully represent the knowledge of the three LLMs in urology subfields. Lastly, the assessed EA might exhibit bias due to the involvement of only two physicians in the analysis, potentially introducing subjective individual variations.

Conclusions

Despite incorporating EA alongside FA, the error rates were ranging from 16 to 35% on the EBU-ISA questions. The additional score gains did not significantly enhance LLMs' overall accuracy rates. This demonstrates that ChatGPT-3.5, ChatGPT-4, and Bing AI exhibit suboptimal urology knowledge and are unsatisfactory proficiency for educational purposes. Substantial variations in LLM performance exist across urological subtopics. Medicine-specific training of LLMs is necessary before these tools can be confidentially recommended for acquiring urological knowledge and effectively integrated into medical education.

Acknowledgments

We extend our heartfelt gratitude to the Executive Committee of the European Board of Urology (EBU) for their support and for confidentially providing the written examination questions from the In-Service Assessment (ISA) in the year 2022.

Statement of Ethics

Following an ethical consultation by the Ethics Committee of the University of Regensburg (Germany) and discussions with the EBU Board, it was determined that no ethical approval was required for this study. The research does not involve any examination of patients or animals.

Conflict of Interest Statement

The authors have nothing to disclose.

Funding Sources

No funding was received for the preparation of this study.

Author Contributions

M.M, L.K., K.E., and K.K.-R. developed study conception and design. Material preparation, data collection, and analysis were performed by K.K.-R., L.K., K.E., and M.M. The first draft of the manuscript was written by L.K., K.E., K.K.-R., and M.M., and the final language and content revision of the manuscript were conducted by S.D.B.-M., M.M., M.R., and M.B. commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability Statement

Data are available upon request from Matthias May, with access subject to restrictions due to privacy laws in Germany.

References

1 Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data. 2021;8(1):53. doi: 10.1186/s40537-021-00444-8.
2 Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Physical Syst. 2023;3:121–54. doi: 10.1016/j.iotcps.2023.04.003.
3 May M, Körner-Riffard K, Marszalek M, Eredics K. Would Uro_Chat, a newly developed generative artificial intelligence large language model, have successfully passed the In-Service Assessment Questions of the European Board of Urology in 2022? Eur Urol Oncol. 2024;7(1):155–6. doi: 10.1016/j.euo.2023.08.013.
4 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. 2024;42(1):20. doi: 10.1007/s00345-023-04749-6.
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. 2023;2(2):e0000198. doi: 10.1371/journal.pdig.0000198.
6 Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the specialty certificate examination in dermatology. Clin Exp Dermatol. 2023:llad255. doi: 10.1093/ced/llad255.
7 Oh N, Choi G-S, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res. 2023;104(5):269–73. doi: 10.4174/astr.2023.104.5.269.
8 Deebel NA, Terlecki R. ChatGPT performance on the American urological association self-assessment study program and the potential influence of artificial intelligence in urologic training. Urology. 2023;177:29–33. doi: 10.1016/j.urology.2023.05.010.
9 Caglar U, Yildiz O, Meric A, Ayranci A, Gelmis M, Sarilar O, et al. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. J Pediatr Urol. 2024;20(1):26.e1–5. doi: 10.1016/j.jpurol.2023.08.003.
10 Deebel NA, Terlecki R. ChatGPT performance on the American Urological Association Self-assessment study program and the potential influence of artificial intelligence in urologic training. Urology. 2023;177:29–33. doi: 10.1016/j.urology.2023.05.010.
11 Huynh LM, Bonebrake BT, Schultis K, Quach A, Deibert CM. New artificial intelligence ChatGPT performs poorly on the 2022 self-assessment study program for urology. Urol Pract. 2023;10(4):409–15. doi: 10.1097/UPJ.0000000000000406.
12 Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in Ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci. 2023;3(4):100324. doi: 10.1016/j.xops.2023.100324.
13 Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? Med Edu Online. 2023;28(1):2220920. doi: 10.1080/10872981.2023.2220920.
14 Suchman K, Garg S, Trindade AJ. Chat generative pretrained transformer fails the multiple-choice American college of gastroenterology self-assessment test. Am J Gastroenterol. 2023;118(12):2280–2. doi: 10.14309/ajg.0000000000002320.
15 Cakir H, Caglar U, Yildiz O, Meric A, Ayranci A, Ozgor F. Evaluating the performance of ChatGPT in answering questions related to urolithiasis. Int Urol Nephrol. 2024;56(1):17–21. doi: 10.1007/s11255-023-03773-0.
16 Musheyev D, Pan A, Loeb S, Kabarriti AE. How well do artificial intelligence chatbots respond to the top search queries about urological malignancies? Eur Urol. 2024;85(1):13–6. doi: 10.1016/j.eururo.2023.07.004.
17 Teoh JY-C, Cacciamani GE, Gomez Rivas J. Social media and misinformation in urology: what can be done? BJU Int. 2021;128(4):397. doi: 10.1111/bju.15517.
18 Klang E, Portugez S, Gross R, R KL, A B, M G, et al. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. BMC Med Educ. 2023;23(1):772. doi: 10.1186/s12909-023-04752-w.
19 Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One. 2023;18(8):e0290691. doi: 10.1371/journal.pone.0290691.
20 Eppler M, Ganjavi C, Ramacciotti LS, Piazza P, Rodler S, Checcucci E, et al. Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology. Eur Urol. 2024;85(2):146–53. doi: 10.1016/j.eururo.2023.10.014.
21 Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine. 2023;90:104512. doi: 10.1016/j.ebiom.2023.104512.