Annual updates of the European Association of Urology — European Society for Pediatric Urology (EAU-ESPU) paediatric urology guidelines: Are large-language models (LLM) better than the usual structured methodology?

L.A. ʻt Hoen [a,*,1], A. van Uitert [b,1], M. Bussmann [c], C. Bezuidenhout [d], M. Ribal [e], S. Canfield [f], Y. Yuan [g,h], M.I. Omar [i], M. Castagnetti [j,k], B. Burgu [l], F. O'Kelly [m], J. Quaedackers [n], Y. Rawashdeh [o], S. Silay [p], A. Bujons [q], G. Bogaert [r], N. Pakkasjarvi [s,t], M. Skott [o], U. Kennedy [u], M. Gnech [v], C. Radmayr [w]

* Correspondence to: Lisette ʻt Hoen, Department of Pediatric Urology, Erasmus Medical Center, Rotterdam, the Netherlands l.thoen@erasmusmc.nl (L.A. ʻt Hoen)

Keywords
Artificial intelligence; Large language models; Pediatric urology; Clinical guidelines

Received 2 December 2024
Revised 28 April 2025
Accepted 28 May 2025
Available online xxx

Summary

Introduction

The European Association for Urology — European Society for Pediatric Urology (EAU-ESPU) guidelines comprise a comprehensive publication of evidence based clinical guidelines for the field of Pediatric urology. The goal is to produce recommendations to optimize patient care and provide an assessment of benefits and harms and possible alternative treatment options. Artificial intelligence (AI) has immensely evolved and is often used in urology. With the emergence of Chat Generative Pre-trained Transformer (ChatGPT) and CoPilot, a new dimension in AI was reached and more widespread use of AI became possible. ChatGPT and CoPilot are both large language models (LLMs).

Objectives

The aim of the current study was to test the ability of LLMs to provide a trustworthy update of two of the chapters of the EAU-ESPU Pediatric Urology Guideline.

Study design

Three LLM's (Chat-GPT 3.5, Chat-GPT 4.0 and CoPilot) were asked to perform a systematic update of the hydrocele and varicocele chapters. For both chapters two standard conversations were written; one humane dialogue and one conversation in which we included minor prompt engineering, i.e. few-shot prompting. All conversations were performed five times by an independent researcher and outcomes were scored for accuracy, consistency and reliability, using several predefined criteria by two reviewers.

Results

A total of sixty conversations were analyzed. All three LLMs were unable to update the guidelines with the recent relevant literature because of the lack of access to the correct scientific databases. Furthermore, a high variability was seen in the responses provided by the LLMs, although the input text was similar every time. The use of basic prompting in the structured conversations compared to the humane responses improved the consistency of the responses. The reproducibility, consistency, and reliability of the updates provided by the LLMs were assessed to be inadequate, despite the use of basic prompting.

Discussion

Development of AI and specific plug-ins for LLMs are developing at a very fast pace. A specific follow-up

[a] Department of Pediatric Urology, Erasmus Medical Center, Rotterdam, the Netherlands [b] Department of Urology, Radboud University Medical Centre, Nijmegen, the Netherlands [c] M. Bussman, Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany [d] Guidelines Office, European Association of Urology, Arnhem, the Netherlands [e] Uro-Oncology Unit, Hospital Clinic, University of Barcelona, Barcelona, Spain [f] Division of Urology, University of Texas McGovern Medical School, Houston, TX, USA [g] Department of Medicine, London Health Science Centre, London, Ontario, Canada [h] Department of Medicine, McMaster University, Hamilton, Ontario, Canada [i] Academic Urology Unit, University of Aberdeen, Aberdeen, UK [j] Department of Surgical, Oncological and Gastroenterological Sciences, University of Padova, Padua, Italy [k] Pediatric Urology Unit, Bambino Gesù Children Hospital and Research Center, Rome, Italy [l] Department of Pediatric Urology, Ankara University School of Medicine, Ankara, Turkey [m] Division of Paediatric Urology, Beacon Hospital Dublin & University College Dublin, Ireland [n] Department of Urology, University Medical Center Groningen, Groningen, the Netherlands [o] Department of Urology, Section of Pediatric Urology, Aarhus, Denmark [p] Division of Pediatric Urology, Department of Urology, Biruni University, Istanbul, Turkey [q] Division of Pediatric Urology, Urology Department Fundació Puigvert Universitat Autonòma Barcelona, Spain [r] Department of Urology, University of Leuven, Leuven, Belgium [s] Dept of Pediatric Surgery, New Children's Hospital, Helsinki University Hospital, Helsinki, Finland [t] Finland Dept of Pediatric Surgery, Section of Urology, University Children's Hospital, Uppsala, Sweden [u] Department of Pediatric Urology, University Children's Hospital Zurich, Zurich, Switzerland [v] Department of Paediatric Urology, Fondazione IRCCS Ca' Granda, Ospedale Maggiore Policlinico, Milan, Italy [w] Pediatric Urology, Medical University of Innsbruck, Innsbruck, Austria [1] Lisette ʻt Hoen and Allon van Uitert share co-first authorship.

https://doi.org/10.1016/j.jpurol.2025.05.030
1477-5131/© 2025 The Authors. Published by Elsevier Ltd on behalf of Journal of Pediatric Urology Company. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

project would be to create specific plug-ins and advanced prompt engineering in cooperation with AI experts for existing LLMs to update the guidelines with access to the relevant databases and correct instructions to follow the handbook of the guidelines.

Conclusion

At the moment LLMs cannot replace the panel members of the EAU Guidelines panel in their work to update the clinical guidelines. They have demonstrated inadequate consistency, reliability, accuracy, and are not able to incorporate new literature.

Introduction

The European Association for Urology (EAU) guidelines comprise a comprehensive publication of evidence based clinical guidelines for the different subspecialties within urology [1]. The goal is to produce recommendations to optimize patient care and provide an assessment of benefits and harms and possible alternative treatment options. These panels are supported by methods experts and some panels also have patient representative as members. The updates are performed in a standardized and reproducible manner to produce clinical guidelines based on best available unbiased evidence [2]. An important part of the EAU guidelines updates is based on systematic reviews that are conducted to synthesize the evidence available [3]. The goal is to deliver clinical practice guidelines that are trustworthy. Therefore, it is important that a robust and up-to-date scope of the available literature is performed in a transparent manner. Also, the quality of evidence and strengths of recommendations should be explicitly stated. Finally, the guidelines should be revised periodically and reconsidered when new evidence becomes available. The current EAU — European Society for Pediatric Urology (ESPU) Paediatric Urology guidelines consist of 23 chapters [4]. A cycle of five years is followed to update the entire guideline, resulting in a yearly update of around five chapters.

In the field of artificial intelligence (AI) robust datasets are combined with computer science to facilitate problem-solving. Deep learning and machine learning are part of AI, and these algorithms can be used for predictions and classifications [5]. In recent years, AI has rapidly evolved, causing a public debate on its ethical use [6]. The use of AI in urology includes prediction models, description of pathology specimens and patient education [7—10]. With the emergence of Chat Generative Pre-trained Transformer (ChatGPT), a new dimension in AI was reached and more widespread use of AI became possible [11]. ChatGPT is a large language model (LLM) which uses deep learning techniques to generate convincing sentences from a huge database of text and data collected during its training phase from the Internet. This makes it able to understand, summarize, generate and predict new content [11].

As the EAU-ESPU Paediatric Urology guidelines panel, we wanted to explore the ability of LLMs to provide a trustworthy update of the guidelines for clinical experts without extensive AI experience. Using ChatGPT versions 3.5, 4.0 (OpenAI, San Francisco, CA, USA) and Copilot (Microsoft, Redmond, Washington, USA), we wanted to identify the possibilities and pitfalls encountered when using AI for this specific purpose.

Our hypothesis was that LLMs might be able to provide support to lighten, ease and accelerate the work, but not replace the added value of the expert panel members.

Material and methods

Two chapters that were previously updated in accordance with the systematic approach described in the handbook for guidelines development were chosen to update using LLMs [2]. The varicocele chapter was published in the EAU-ESPU guidelines on Paediatric Urology 2023 version [4] and the hydrocele chapter was published in the EAU-ESPU guidelines on Paediatric Urology 2024 version [12]. These two chapters were chosen for two reasons. First, because they were updated recently, which enabled us to test if the LLMs were able to find recent literature. Second, since we expected that this would be a challenging task for the LLMs, these chapters were chosen because they were small and compact, without controversial or quickly developing topics. Although a paragraph was added to the hydrocele chapter in the 2024 update regarding abdominoscrotal hydrocele, this was not included in the requested update from the LLMs to keep the task straightforward.

ChatGPT was chosen because it is the largest and most well-known language model available for the public. We used the free version (ChatGPT3.5) as well as the paid version (ChatGPT4.0). Furthermore, we chose to include Copilot as an additional language model, since this model has internet access in contrary to ChatGPT and is gaining popularity. The study protocol was registered at Open Science Framework (OSF) on April 15, 2024 (https://osf.io/u9rtp/).

Input source

For both the hydrocele and varicocele chapter two standard conversations were written: one humane dialogue (Supplement 1), which was a normal conversation with the LLM without prompt engineering, and one structured dialogue (Supplement 2), which included minor prompt engineering, i.e. few-shot prompting. Together with an AI model expert, we chose to provide a task description as we would for new guidelines panel members. We were aware of the implicit assumptions and jargon that were used in the prompts, since this is a very difficult task for the AI model to perform.

The models were asked to perform the following tasks:

1. Search for new evidence using these scientific medical databases (PubMed, Ovid, Embase and Cochrane Central Register of Controlled Trials and the Cochrane Database of Systematic Reviews)
2. Extract articles and abstracts
3. Select only articles and abstracts from the years 2017—2022
4. Select those that provide new evidence
5. Make an updated version of the chapter based on the new evidence

6. Show a summary of changes at the end
7. Provide the references or locations where the new evidence was found
8. Compare these findings with the old chapter and report what is new
9. Compose recommendations based upon evidence-based methodology which can be found in the handbook.

Additionally, the models were asked to answer a clinical scenario question using the provided clinical guideline.

Each conversation was repeated 5 times per chat to ensure consistency and reliability and each individual conversation was introduced to ChatGPT-3.5, ChatGPT-4.0 and Copilot via a new tub. The conversations were performed consecutively by an independent team member who was not involved with the development of the protocol or interpretation of the output of the AI models (C.B.).

Output interpretation

The output from the 3 language models was compared to check for accuracy, consistency and reliability. The predefined criteria that were used are: ability to access EAU Guideline Handbook, ability to access reliable sources (external scientific databases), provides an update that is scientifically correct, provides scientifically correct recommendations, provides scientifically correct references and provides a correct answer to the clinical scenario. These criteria were scored as yes/no to determine the variation in answers provided by the AI models. For the hydrocele and varicocele chapter two criteria for each chapter were used to score if the response provided by the LLMs was scientifically correct. For the hydrocele chapter: no standard use of ultrasound, and early surgery is recommended when there is a suspicion of a concomitant inguinal hernia or underlying testicular pathology. For the varicocele chapter: in case of right sided varicocele an ultrasound of the retroperitoneum is recommended, and if the surgical indications were correctly described.

All conversations were independently scored to evaluate the abovementioned predefined criteria by two paediatric urologists from the EAU-ESPU Paediatric Urology guidelines panel (AvU and LtH). Interrater reliability was evaluated by calculating a percent agreement between raters per chapter and by calculating Cohen's kappa for the total of measurements [13].

Results

Between April 16th 2024, and April 19th 2024, all separate conversations were performed five times with the three different LLMs using both the humane and structured dialogue, resulting in a total of 60 conversations (see Table 1). Two sample conversations are shown in Supplement 3 and 4.

Although the conversations were all performed by one person in an identical fashion, the answers and performance from the LLMs were different every time and often contradictory. For example, in 2 out of 5 conversations using the humane hydrocele dialogue, Chat-GPT3.5 answered that it was able to access external scientific websites, however, in 3 out of 5 it could not.

Another example was seen in the structured conversations. In 1 out of 5 conversations using the structured dialogue, ChatGPT3.5 falsified very realistic appearing references of non-existent scientific articles to support the updated findings: *Smith, J.* et al. *"Long-term outcomes of varicocele repair in adolescents: a systematic review and meta-analysis." Journal of Pediatric Urology, 2020*. After checking the journal, this study was non-existent.

Large language models (LLM)

When comparing the results of the three different LLMs; ChatGPT3.5, ChatGPT4.0 and CoPilot, several differences were seen (see Table 1).

First, the ability to access external websites was different; Chat-GPT3.5 reported to be unable to access the EAU guideline development handbook in the humane conversations, while Chat-GPT4.0 and Copilot did not. Also, the ability to access scientific websites (PubMed, Ovid, Embase, or Cochrane) was different between the LLMs, both in the humane and structured dialogues.

When asked to provide an updated version of the chapter, Chat-GPT3.5 was able to provide this in most conversations, but only 5 out of 20 conversations were scientifically correct. However, no new information was incorporated in these correct updates, it only provided a scientifically correct summary without information loss.

Chat-GPT4.0 and Copilot were not able to write a scientifically correct update. Chat-GPT4.0 provided tools to assist in the update process (12/20 conversations) or made a summary of the chapter (4/20 conversations). Although Copilot was the only LLM that was able to find new real references, it could not adequately incorporate this in an updated chapter. Furthermore, this was not done using a systematic methodology, and the articles provided were often old and outside the scope of the chapter.

Conversation types

When comparing the humane versus the structured dialogues, several differences were seen. In the humane dialogue ChatGPT3.5 was able to read the EAU guideline development handbook in 1 out of 10 conversations, while in the structured dialogue it was able to in 9 out of 10 conversations. Similar differences were seen regarding access of external scientific websites (4/10 versus 10/10).

Hydrocele versus varicocele

When comparing the different subchapters used for this hypothetical update, again a variability was seen between all LLMs and conversation types. The clinical scenario in the varicocele chapter proved to be more difficult to answer correctly when compared to the hydrocele scenario (21/30 correct answers versus 30/30 correct answers, respectively).

Table 1a Dialogue results per large language model for the hydrocele chapter

	Chat-GPT3.5	Chat-GPT4	Copilot
Hydrocele — humane dialogue			
Is able to read EAU Guideline development handbook	00000	11011	11111
Is able to access external websites (Pubmed, Ovid, Embase, or Cochrane library)	00110	10000	00000
Provides an updated version of the chapter	01111	10010	01111
- If so, this update is scientifically correct	x0000	0xx0x	x0000
- If so, provides a summary of the chapter	x1111	1xx1x	x1111
- If so, only makes a hypothetical update	x0000	0xx0x	x0000
- If not, provides tools to assist in update process	1xxxx	x11x1	1xxxx
Provides recommendations from the literature review	11111	10011	00010
- If so, this update is scientifically correct	11100	0xx00	xxx0x
Provides a summary of changes made to the chapter from the literature review	11111	10010	01111
- If so, this update is scientifically correct	00000	0xx0x	x0000
Provides new real references used for the update	00000	00000	11111
Fabricates fake references used for the update	00000	00000	00000
Provides a correct answer to clinical scenario 1	11111	11111	11111
Hydrocele — structured dialogue			
Is able to read EAU Guideline development handbook	11111	11111	11111
Is able to access external websites (Pubmed, Ovid, Embase, or Cochrane library)	11111	00000	11110
Provides an updated version of the chapter	01011	00000	00001
- If so, this update is scientifically correct	x1x00	xxxxx	xxxx0
- If so, provides a summary of the chapter	x1x11	xxxxx	xxxx1
- If so, only makes a hypothetical update	x0x00	xxxxx	xxxx0
- If not, provides tools to assist in update process	0x0xx	11111	0000x
Provides recommendations from the literature review	11111	00000	11111
If so, this update is scientifically correct	11111	xxxxx	00000
Provides a summary of changes made to the chapter from the literature review	00001	00000	11111
- If so, this update is scientifically correct	xxxx0	xxxxx	00000
Provides new real references used for the update	00000	00000	11111
Fabricates fake references used for the update	00001	00000	00000
Provides a correct answer to clinical scenario 1	11111	11111	11111

Model conversations were performed five separate times in all three databases.
0 = no.
1 = yes.
x = not applicable.

Criteria for accuracy, consistency and reliability

When comparing the predefined criteria for accuracy, consistency and reliability, the structured conversation with Chat-GPT3.5 has the highest cumulative score (21/30 for the hydrocele chapter, 20/30 for the varicocele chapter), followed by the structured conversations with Copilot (19/30 for the hydrocele chapter, 18/30 for the varicocele chapter), see Table 2a and Table 2b.

Interrater reliability

For the hydrocele chapter the two researchers (AvU and LtH) scored 407 of the total 420 predefined criteria in agreement, with an interrater reliability of 97.0 %. For the varicocele chapter the two researchers (AvU and LtH) scored 403 out of the total 420 predefined criteria in agreement, with an interrater reliability of 96.0 %. The combined interrater reliability, calculated using Cohen's kappa, was 0.94 (95 % confidence interval 0.92 to 0.96), which is an almost perfect agreement.

Discussion

The strength of the EAU guidelines is the standardized methodology and evidence-based approach which is used to create and update the guidelines on a regular basis. The EAU guidelines must be accurate, consistent, reliable and up to date, since they are widely used by clinicians.

Our hypothesis that LLMs cannot replace the EAU-ESPU guidelines panel members has been confirmed when considering these important characteristics. LLMs have been unable to perform one of the most essential steps of the guidelines panel, which is updating the guidelines with the recent relevant literature because of the lack of access to scientific medical databases, such as PubMed, Ovid, Embase and Cochrane Central Register of Controlled Trials and the Cochrane Database of Systematic Reviews. When an update was provided using real references by scoping the relevant databases, either the program just copied the old references as new, or other references were found which were outside the scope of the chapter text (i.e. case reports, adult papers). Therefore, the tested LLMs cannot

Please cite this article as: 't Hoen LA et al., Annual updates of the European Association of Urology — European Society for Pediatric Urology (EAU-ESPU) paediatric urology guidelines: Are large-language models (LLM) better than the usual structured methodology?, Journal of Pediatric Urology, https://doi.org/10.1016/j.jpurol.2025.05.030

Table 1b Dialogue results per large language model for the varicocele chapter

	Chat-GPT3.5	Chat-GPT4	Copilot
Varicocele — humane dialogue			
Is able to read EAU Guideline development handbook	00100	11111	11111
Is able to access external websites (Pubmed, Ovid, Embase, or Cochrane library)	01010	00000	00000
Provides an updated version of the chapter	01010	00101	00000
- If so, this update is scientifically correct	x1x0x	xx0x0	xxxxx
- If so, provides a summary of the chapter	x1x1x	xx1x1	xxxxx
- If so, only makes a hypothetical update	x0x0x	xx0x0	xxxxx
- If not, provides tools to assist in update process	1 × 1 × 1	00x0x	10110
Provides recommendations from the literature review	11111	11111	01000
- If so, this update is scientifically correct	00110	00000	x0xxx
Provides a summary of changes made to the chapter from the literature review	01010	00101	00000
- If so, this update is scientifically correct	x0x0x	xx0x0	xxxxx
Provides new real references used for the update	00000	00000	11111
Fabricates fake references used for the update	00000	00000	00000
Provides a correct answer to clinical scenario 1	11110	11111	0000x
Varicocele — structured dialogue			
Is able to read EAU Guideline development handbook	11110	11111	11111
Is able to access external websites (Pubmed, Ovid, Embase, or Cochrane library)	11111	10000	11111
Provides an updated version of the chapter	11111	10000	10001
- If so, this update is scientifically correct	01110	0xxxx	0xxx0
- If so, provides a summary of the chapter	11111	0xxxx	0xxx0
- If so, only makes a hypothetical update	00000	1xxxx	0xxx0
- If not, provides tools to assist in update process	xxxxx	x1111	x010x
Provides recommendations from the literature review	11111	10000	11011
If so, this update is scientifically correct	10111	0xxxx	00x00
Provides a summary of changes made to the chapter from the literature review	11111	10000	11001
- If so, this update is scientifically correct	00000	0xxxx	00xx0
Provides new real references used for the update	00000	00000	11111
Fabricates fake references used for the update	00001	00000	00000
Provides a correct answer to clinical scenario 1	01111	11111	10110

Model conversations were performed five separate times in all three databases.
1 = Yes.
0 = No.
x = not applicable.

assist or replace the panel members with the literature searches or interpretations.

Furthermore, a high variability was seen in the responses provided by the LLMs, although the input text was similar every time. This was not only seen between the different LLMs, but also within the same LLM. The responses ranged from different content to showing contrasting abilities in terms of access to databases or abilities to perform an update of the chapter text. The use of basic prompting in the structured conversations compared to the humane responses did improve the consistency of the responses. However, the reproducibility, consistency, and reliability of the updates provided by the tested LLMs remained inadequate.

Another aspect of the high quality of the EAU guidelines is the accuracy of the information. To make accurate and trustworthy guidelines it is important to ensure the use of reliable resources. This characteristic needs to be highlighted, because the different LLMs demonstrated important variabilities in reproducing resources for their updates. ChatGPT3.5 could not provide references in most of the conversations, and when it provided references in two conversations, these were fake. However, these references looked very convincing, since the authors described in the fake references had previous publications on the topic of interest, and the journals mentioned in the references are existing journals in the field of urology. This falsification of references by ChatGPT has previously been described [14]. This is a topic which should warrant much caution, since these references can make the provided data feel more powerful, although they are fake.

An additional aspect of interest was the summary of changes the LLMs provided. Often the summary of changes did not correlate to the updated chapter text it provided, e.g. information was mentioned in the summary of changes, but these facts were not part of the updated text. Also, several times a summary of changes was made without the presence of an updated chapter text.

When making an objective interpretation of output of the LLMs, there was a difference between the LLMs and their ability to access the internet. It is important to note that ChatGPT3.5 is not able to access external sources or browse the internet real-time. The knowledge it has is based on the databases with which it was created (2021) and is not updated automatically. Interestingly, ChatGPT4.0 and

Table 2a Criteria for accuracy, consistency and reliability - Hydrocele.

	ChatGPT 3.5		**ChatGPT 4.0**		**CoPilot**	
	Humane	Structured	Humane	Structured	Humane	Structured
Access to EAU guidelines book	(0/5)	(5/5)	(4/5)	(5/5)	(5/5)	(5/5)
Access to external databases	(2/5)	(5/5)	(1/5)	(0/5)	(0/5)	(4/5)
Scientifically correct update	(0/5)	(1/5)	(0/5)	(0/5)	(0/5)	(0/5)
Scientifically correct recommendations	(3/5)	(5/5)	(0/5)	(0/5)	(0/5)	(0/5)
Scientifically correct references	(0/5)	(0/5)	(0/5)	(0/5)	(5/5)	(5/5)
Correct answer clinical scenario	(5/5)	(5/5)	(5/5)	(5/5)	(5/5)	(5/5)
Total	**10/30**	**21/30**	**10/30**	**10/30**	**15/30**	**19/30**

Table 2b Criteria for accuracy, consistency and reliability - Varicocele.

	ChatGPT 3.5		**ChatGPT 4.0**		**CoPilot**	
	Humane	Structured	Humane	Structured	Humane	Structured
Access to EAU guidelines book	(1/5)	(4/5)	(5/5)	(5/5)	(5/5)	(5/5)
Access to external databases	(2/5)	(5/5)	(0/5)	(1/5)	(0/5)	(5/5)
Scientifically correct update	(1/5)	(3/5)	(0/5)	(0/5)	(0/5)	(0/5)
Scientifically correct recommendations	(2/5)	(4/5)	(0/5)	(0/5)	(0/5)	(0/5)
Scientifically correct references	(0/5)	(0/5)	(0/5)	(0/5)	(5/5)	(5/5)
Correct answer clinical scenario	(4/5)	(4/5)	(5/5)	(5/5)	(0/5)	(3/5)
Total	**10/30**	**20/30**	**10/30**	**11/30**	**10/30**	**18/30**

Copilot both claim to have access to the internet. During the conversations they were almost always able to access the handbook (19/20). ChatGPT4.0 was unable to access the relevant databases (e.g. PubMed) in 9/10 of the humane conversations and 9/10 of the structured conversations. Copilot was unable to access the relevant databases in 10/10 of the humane conversations but was only unable to access these databases in 1/10 of the structured conversations. The LLMs caused confusion about their ability to access the relevant databases in multiple conversations. ChatGPT3.5 and Copilot stated they would access the relevant databases and consequently provide a chapter update when asked to do so. However, only when asked on what references the update was based did the LLM state that it did not have access to these specific databases, thus contradicting its earlier statement. An example of this is seen in Supplement 3. The variation in responses makes the reliability of the programs inadequate for chapter updates.

We have used two types of conversations for testing the current hypothesis: humane and structured conversation, which includes basic prompt engineering. The decision to include a humane conversation is that this is the way most people, including paediatric urologists, use the different LLMs in daily life. We wanted to explore if adding basic prompt engineering (structured conversations), which is still something that the common urologist would be able to perform, provided better results. After the structured conversations, responses provided by the different LLMs became more consistent and the scientific quality improved, especially for ChatGPT3.5 and Copilot.

When we make a head-to-head comparison between the different LLMs, a high variability exists in all LLMs. Copilot seems to be the most consistent LLM in its output and can locate real scientific references, although the relevance of these studies to the actual chapter was low. ChatGPT3.5 is the most used LLM in daily practice and we want to emphasize caution with its use given the fact that ChatGPT3.5 provided false non-existent references. However, basic prompt engineering clearly improved the outcomes of ChatGPT3.5.

We have to conclude that at this moment, current LLMs are inadequate to update the EAU Paediatric Urology guidelines. However, there could be different roles for the LLMs, now and in the future. Firstly, the different LLMs can be used to summarize the updated chapters scientifically correctly, which could save time for the EAU Guidelines Pocket Handbook or the EAU guidelines app. However, a word-by-word validation of the summary needs to be performed by the panel members afterwards, to ensure a correct outcome.

Secondly, LLMs could be used to improve the quality of the language used in the different guidelines, since guidelines panels are composed of urologists with different backgrounds, and most are not native English speakers. This is currently a very time-consuming task and can be done much quicker when the first corrections have been made by LLMs. However, again a word-by-word analysis should be performed of the updated text afterwards to prevent loss of relevant information and additions of wrongful information. Lastly, the programs could be used to formulate multiple choice questions based on provided text. This could assist the Panel members as well, since they are consulted yearly to provide these questions for the annual urology exams.

Limitations and future implications

A limitation of this study is that the updates provided by the different LLMs are constrained by the information accessible to these models. ChatGPT3.5 was the first AI program that has been widely available and free of charge. However, as stated earlier it is a language-based model. Development of AI and specific plug-ins for LLMs are developing at a very fast pace. A specific follow-up project would be to create specific plug-ins and advanced prompt engineering in cooperation with AI experts for existing LLMs to update the guidelines, incorporating access to the relevant databases and correct instructions to follow the handbook of the guidelines [2].

Conclusion

Current LLMs cannot replace the panel members of the EAU-ESPU Guidelines panel in their work to update the clinical guidelines. They have demonstrated inadequate consistency, reliability, accurateness, and are not able to incorporate new literature. AI is a quickly evolving field and new options are being presented with a high frequency. When specific plug-ins are developed for updating the EAU guidelines, following the stringent standardized methodology and access to the relevant databases, LLMs might be used to assist the work of the panel members in the future. At this moment, LLMs could be used to improve the quality of the English language used in the guideline or update the pocket guidelines. However, the outcomes need to be checked. Caution is warranted when the LLMs provide references, since these could be non-existent and not based on real study data, which can be misleading and dangerous.

Funding source

The EAU has paid for the use of Chat GPT 4.0. They had no involvement in the study design or data interpretation.

Conflict of interest

No conflict of interest.

References

[1] EAU Guidelines. Edn. presented at the EAU annual congress Milan. 2023. ISBN 978-94-92671-19-6.
[2] European Association of Urology. Guidelines office development handbook. Arnhem, The Netherlands: EAU; 2023. https://uroweb.org/eau-guidelines/methodology-policies.
[3] Knoll T, Omar MI, MacLennan S, Hernandez V, Canfield S. Yuan Y.,et al. Key steps in conducting systematic reviews for underpinning clinical practice guidelines: methodology of the European association of urology. Eur Urol 2018;73:290—300. https://doi.org/10.1016/j.eururo.2017.08.016.
[4] Radmayr C, Bogaert G, Burgu B, Castagnetti M, Dogan H, O Kelly F, et al. EAU guidelines on paediatric urology 2023. Edn Presented at the EAU annual congress Milan 2023.
[5] Russell SJ, Norvig P. Artificial intelligence: a modern approach. 4th ed. Boston: Pearson; 2020. ISBN 9780134610993.
[6] Corsello A, Santangelo A. May artificial intelligence influence future pediatric research? -The case of ChatGPT. Children 2023 Apr 21;10(4):757. https://doi.org/10.3390/children10040757. PMID: 37190006; PMCID: PMC10136583.
[7] Wu S, Hong G, Xu A, Zeng H, Chen X, Wang Y, et al. Artificial intelligence-based model for lymph node metastases detection on whole slide images in bladder cancer: a retrospective, multicentre, diagnostic study. Lancet Oncol 2023 Apr;24(4):360—70. https://doi.org/10.1016/S1470-2045(23)00061-X. Epub 2023 Mar 6. PMID: 36893772.
[8] Suarez-Ibarrola R, Sigle A, Eklund M, Eberli D, Miernik A, Benndorf M, et al. Artificial intelligence in magnetic resonance imaging-based prostate cancer diagnosis: where do we stand in 2021? Eur Urol Focus 2022 Mar;8(2):409—17. https://doi.org/10.1016/j.euf.2021.03.020. Epub 2021 Mar 25. PMID: 33773964.
[9] de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. npj Digit Med 2022 Jan 10; 5(1):2. https://doi.org/10.1038/s41746-021-00549-7. PMID: 35013569; PMCID: PMC8748878.
[10] Görtz M, Baumgärtner K, Schmid T, Muschko M, Woessner P, Gerlach A, et al. An artificial intelligence-based chatbot for prostate cancer education: design and patient evaluation study. Digit Health 2023 May 2;9:20552076231173304. https://doi.org/10.1177/20552076231173304. PMID: 37152238; PMCID: PMC10159259.
[11] https://chat.openai.com/chat.
[12] Radmayr C, Bogaert G, Bujons A, Burgu B, Castagnetti M. t Hoen L.,et al. EAU guidelines on paediatric urology. In: Presented at the EAU annual congress Paris; 2024.
[13] Landis JR, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159—74.
[14] Wu RT, Dang RR. ChatGPT in head and neck scientific writing: a precautionary anecdote. Am J Otolaryngol 2023 Jul 6;44(6): 103980. https://doi.org/10.1016/j.amjoto.2023.103980. Epub ahead of print. PMID: 37459740.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpurol.2025.05.030.

Please cite this article as: 't Hoen LA et al., Annual updates of the European Association of Urology — European Society for Pediatric Urology (EAU-ESPU) paediatric urology guidelines: Are large-language models (LLM) better than the usual structured methodology?, Journal of Pediatric Urology, https://doi.org/10.1016/j.jpurol.2025.05.030