European Journal of
Human Health

ORIGINAL ARTICLE

ASSESSING THE ACCURACY OF AI LANGUAGE MODELS IN PROVIDING INFORMATION ON URINARY INCONTINENCE: A COMPARATIVE STUDY

Burhan COSKUN [1], Omer BAYRAK [2], Gokhan OCAKOGLU [3], Halil Mustafa ACAR [1], Onur KAYGISIZ [1]

[1]Department of Urology, Uludag University Medical Faculty, Bursa, TURKEY
[2] Department of Urology, Gaziantep University Medical Faculty, Bursa, TURKEY
[3]Department of Biostatistics, Uludag University Medical Faculty, Bursa, TURKEY

ABSTRACT

Aim: To assess the accuracy and comprehensiveness of health information generated by different large language models (LLMs) focusing on urinary incontinence.

Methods: Using the website www.answerthepublic.com, we retrieved the most frequently searched questions related to urinary incontinence. After applying exclusion criteria, the chosen questions, categorized into definition/diagnosis, causes, treatment, complications, and others, were input into LLMs: GPT-3.5, GPT-4, and BARD.Outputs were assessed for accuracy and comprehensiveness by two urologists using a Likert scale.

Results: Of the initial 630 questions, 38 were selected for analysis. GPT-4 demonstrated superior performance, with 73.68% of its responses achieving the maximum accuracy score, significantly outperforming GPT-3.5 (42.11%) and BARD (28.95%). In terms of comprehensiveness, GPT-4 also excelled with a score of 71.05%, whereas GPT-3.5 and BARD scored 36.84% and 28.95% respectively.Forthe 'causes' category, GPT-4 provided significantly more comprehensive responses.

Conclusion: While all LLMs generated relevant health information on urinary incontinence, GPT-4 showed superior accuracy and comprehensiveness. However, the potential for generating incorrect information by these models necessitates caution in their utilization.

Keywords: Urinary Incontinence, Large Language Models, Patient Information.

Corresponding Author: Burhan COSKUN burhanc@uludag.edu.tr

Received: ***August 16, 2023;*** *Accepted:* ***August 22, 2023;*** *Published Online:* ***September 01, 2023***

Cite this article as: Coskun, B., Bayrak, O., Ocakoglu, G., Acar, H. M. & Kaygisiz, O.(2023). Assessing the Accuracy of AI Language Models in Providing Information on Urinary Incontinence: A Comparative Study. European Journal of Human Health 3(3),61-70.

INTRODUCTION

Urinary incontinence, a condition defined by the involuntary loss of urine, significantly impacts the quality of individuals across various age groups, regardless of gender[1], [2]. This prevalent condition can be managed through multiple treatment options, including conservative treatments, physiotherapy, medication, and surgical interventions[3]. Despite the variety of treatments available, patients may face challenges communicating their symptoms and condition due to embarrassment or societal stigma [4]. In such scenarios, the Internet often serves as an alternative source of information, allowing patients to explore their treatment options in privacy.

The internet accomplished an exponential growth over the last two decades, becoming a primary source of health information[5]. However, its inherent openness allows for dissemination of incorrect and potentially harmful information, making the accuracy and reliability of online content questionable[6]. In the era of shared decision-making, where patients make informed choices about their treatment in consultation with healthcare professionals, it is becoming increasingly important [7].

Recently, large language models (LLMs), such as those developed by OpenAI (GPT-3.5 and GPT 4), have gained a lot of interest from the public. They are capable of generating human-like text in almost every aspect of life, including healthcare[8]. LLMs are pretrained on billions of inputs and are capable of producing the most relevant continuation in the concept of a statistical machine. However, it should be noted that they are not error-free[9].

The accuracy of patient information is of paramount importance in healthcare. Reliable, clear, and comprehensible health information forms the basis for informed decision-making, patient empowerment, and adherence to treatment plans[10].

Although the field is in its infancy, there is currently no consensus on the performance of LLMs in the context of delivering patient information[11].

In this study, we aim to evaluate of the accuracy, and comprehensibility of health information generated by different LLM models, with a specific focus on urinary incontinence.

METHODS

Data Collection

On May 24, 2023, the term "urinary incontinence" was entered on the website www.answerthepublic.com to collect the most commonly searched questions related to urinary incontinence on Google. Only the results related to “questions” were collected and the results regarding the statements (results of comparisons, related, alphabeticals, duplicates, prepositions) non-English, and irrelevant results were removed. The results were categorized into one of the following categories: definition and diagnosis,

causes, treatment, complications, and other factors.

Each remaining question after applying the exclusion criteria was applied to the following LLM models: GPT3.5, GPT4, and BARD. The outputs from each model were recordedand two urologists separately assessed the accuracy of the models' results.

Data Analysis

The reviewers used a Likert scale, developed by Johnson et al., to gauge the accuracy and completeness of the responses provided by the AI models[12].

The accuracy scale was as follows:

1. Completely incorrect
2. More incorrect than correct
3. Approximately equal correct and incorrect
4. More correct than incorrect
5. Nearly all correct
6. Correct

The completeness scale was structured as follows:

1. Incomplete - addresses some aspects of the question, but significant parts are missing or incomplete
2. Adequate - addresses all aspects of the question and provides the minimum amount of information required to be considered complete
3. Comprehensive - addresses all aspects of the question and provides additional information or context beyond what was expected.

STATISTICAL ANALYSIS

The Fisher-Freeman-Halton test was used in the comparison between language models according to accuracy and comprehensiveness ratios. In the case of general significance, subgroup analyses were performed using Bonferroni correction. Type I error level was accepted as 5% in the analyses performed using the SPSS (IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.) software.

Inter-reviewer Agreement

The inter-reviewer agreement was evaluated using the intraclass correlation coefficient (ICC). ICC values range from 0 to 1, with higher values indicating better agreement between the observers. The ICC value in this study was 0.85, denoting a high level of agreement between the reviewers.

RESULTS

An initial search onwww.answerthepublic.com yielded a total of 630 results pertaining to "urinary incontinence". Following the implementation of the predefined exclusion criteria, 38 relevant questions was finalized for further analysis (Figure 1).

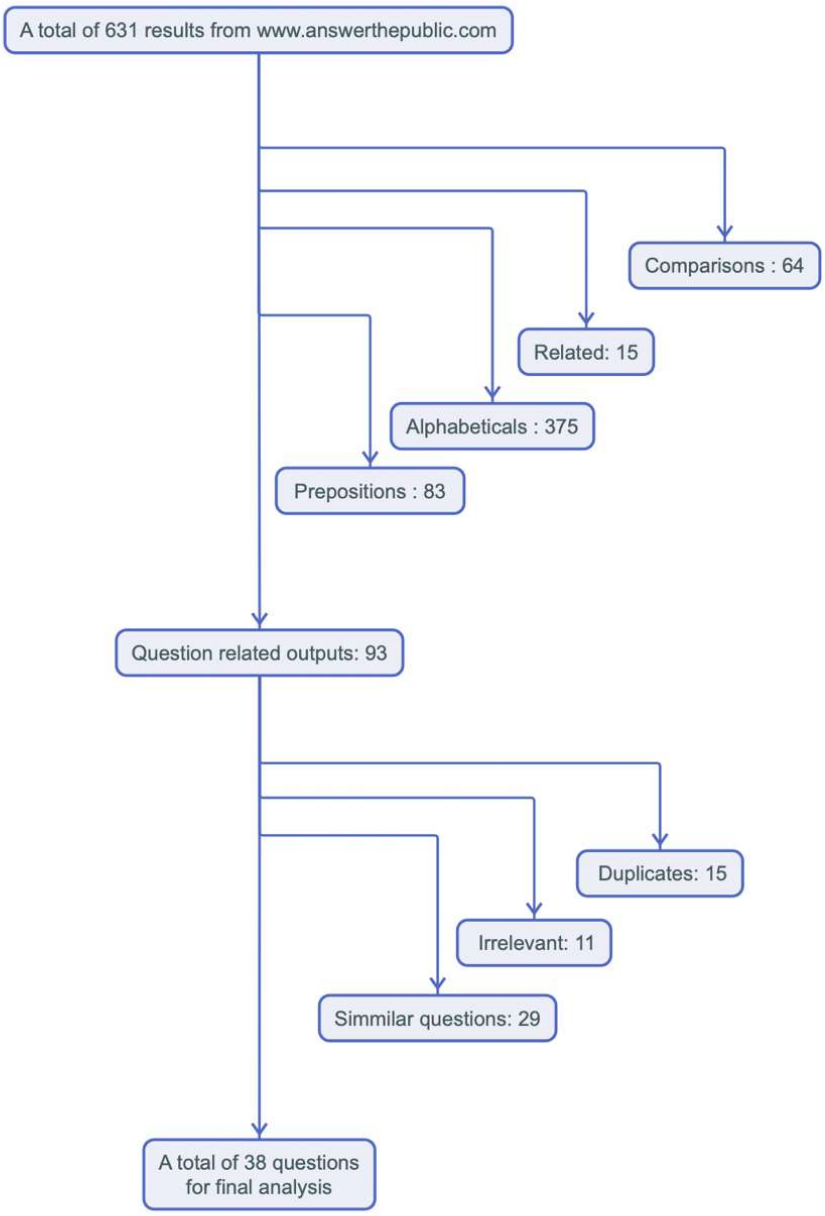

Figure 1: Identifying Eligible Questions from www.answerthepublic.com

In the assessment of accuracy across all questions, the proportion of responses achieving the maximum score was 42,11% for GPT-3.5, 73,68% for GPT-4, and 28,95% for BARD (Figure 2). GPT-4 achieved a significantly higher accuracy rate compared to GPT-3.5 and BARD (p=0.015)On the other hand, the rate of totally incorrect responses was found to be 2,63% for GPT3, 2,63% for GPT4, and 5,26% for BARD.The comparison of accuracy scores for LLM models across question categories was summarized in Table 1.Despite GPT4 demonstrating superior accuracy rates in comparison to the other models, the observed differences were not statistically significant.

Figure 2: Assessment of accuracy across all questions

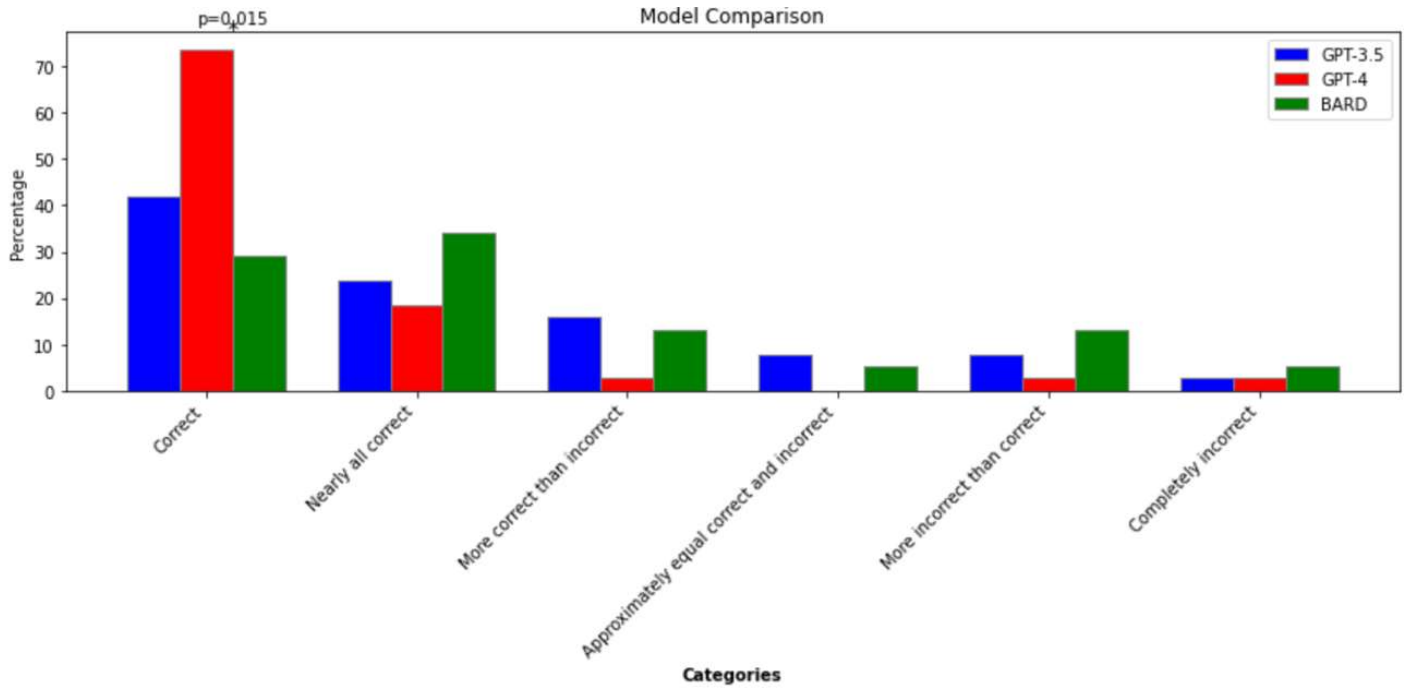

Figure 2: Assessment of accuracy across all questions

Table 1:Comparison of accuracy scores for LLM models across question categories

	GPT 3.5	**GPT 4**	**BARD**	**p-value¥**
Definition and Diagnosis (N=8)				0,496
Correct	62.5%	75%	37.5%	
Nearly all correct	37.5%	25%	50%	
More correct than incorrect	0	0	12.5%	
Causes (N=12)				0.252
Correct	25%	58.3%	8.3%	
Nearly all correct	16.7%	25%	33.3%	
More correct than incorrect	25%	0	8.3%	
Approximately equal correct and incorrect	8.3%	0	16.7%	
More incorrect than correct	16.7%	8.3%	25%	
Completely incorrect	8.3%	8.3%	8.3%	

Treatment (N=6)				0.403
Correct	50%	83.3%	33.3%	
Nearly all correct	16.7%	16.7%	16.7%	
More correct than incorrect	33.3%	0	50.0%	
Complications (N=8)				0.246
Correct	25%	75%	25%	
Nearly all correct	25%	12.5%	37.5%	
More correct than incorrect	12.5%	12.5%	0	
Approximately equal correct and incorrect	25%	0	0	
More incorrect than correct	12.5%	0	25%	
Completely incorrect	0	0	12.5%	
OtherFactors (N=4)				>0,999
Correct	75%	100%	75%	
Nearly all correct	25%		25%	

¥: Fisher-Freeman-Halton test

When evaluating the completeness of responses across all questions, GPT-3.5 achieved a comprehensive output rate of 36.84%, while GPT-4 achieved 71.05%, and BARD achieved 28.95%. GPT-4 had a significantly higher score than the other models (p:0.008) (Figure 3).The rate of entirely incomplete responses was found to be 21,05% for GPT3, 7,89% for GPT4, and 23,68% for BARD.

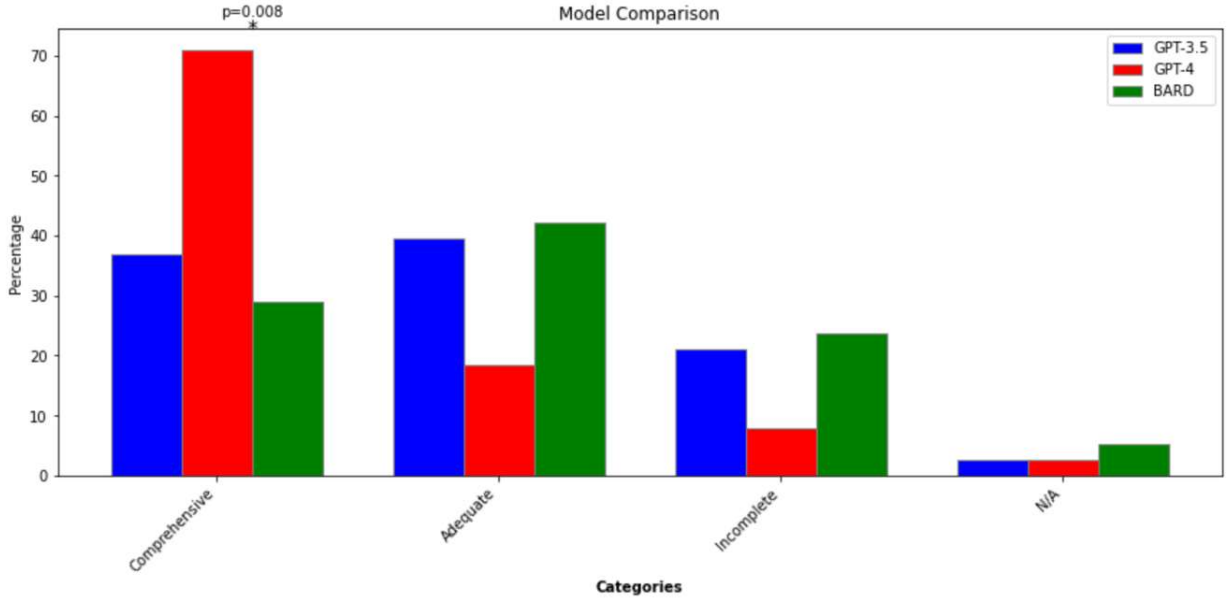

Figure 3: Assessment of completeness across all questions

The comparative analysis of completeness scores among LLM models across the different question categories is presented in Table 2.

Table2 :Comparison of completeness scores for LLM models across question categories

	GPT 3.5	**GPT 4**	**BARD**	**p-value¥**
Definition and Diagnosis (N=8)				0.703
Comprehensive	62.5%	75%	37.5%	
Adequate	25%	12.5%	37.5%	
Incomplete	12.5	12.5	25%	
Causes (N=12)				**0.037**
Comprehensive	16.7% [a]	75% [b]	16.7% [a]	
Adequate	41.7% [a]	8.3% [a]	33.3% [a]	
Incomplete	33.33% [a]	8.3% [a]	41.7% [a]	
N/A	8.33% [a]	8.3% [a]	8.3% [a]	
Treatment (N=6)				0.280
Comprehensive	33.3%	66.7%	33.3%	
Adequate	66.7%	16.7%	66.7%	
Incomplete		16.7%		
Complications (N=8)				0.473
Comprehensive	25%	50%	37.5	
Adequate	37.5%	50%	37.5%	
Incomplete	37.5%		16.7%	
N/A			8.3%	
OtherFactors (N=4)				
Comprehensive	75%	100%	25%	0.200
Adequate	25%		75%	
Incomplete				

¥: Fisher-Freeman-Halton test N/A : Not avaliable

In the 'causes' category, GPT4wassignificantly betterthan its counterparts by providing much more extensive responses (p=0.037). However, when examining the other question categories, although GPT4 showed higher rates of completeness compared to the other models, these differences were not statistically significant.

DISCUSSION

The primary result of this study is the differing levels of accuracy and completeness exhibited by three distinct LLMs. To our knowledge, this is the first study to assess the performance of different LLMs in the context of patient information for urinary incontinence.Our findings highlight the considerable enhancement in accuracy and completeness of responses in GPT-4 compared to its predecessor, GPT-3.5. Additionally, the performance of the GPT-4 model was significantly superior to that of the BARD model.GPT-4 provided significantly more comprehensive responses for the 'causes' category. This can be explained by its higher consciousness compared to the earlier model.

GPT, or Generative Pretrained Transformer, is an LLM developed by OpenAIthat uses the transformer architecture in natural language processing tasks. It is "pretrained" on vast amounts of text data, enabling it to generalize and adapt to more niche requirements. With more sophisticated models, GPT can understand context, generate coherent text over long passages, and address a wide range of natural language processing tasks with unprecedented accuracy[13].

Our group recently reported suboptimal performance of GPT-3.5 as a patient information source for prostate cancer, when compared to a reference material. To evaluate this, we used complex metrics such as precision and recall, implementing true positive, false positive, true negative, and false negative measurements [11].Another study on ChatGPT 3.5 as a patient education tool for robotic-assisted radical prostatectomy showed promising findings. The study used 14 questions from the British Association of Urological Surgeons patient information leaflet, ensuring reliable evaluation. The results showed that ChatGPT had 78.6% agreement with the leaflet's figures and 92.9% of responses were accurate and relevant to potential patient queries, highlighting the tool's reliability in providing information[14].

For LLMs use of spesific prompt can make a difference. In their study for potential use of ChatGPT for information regarding anterior cruciate liagement with specific prompts to medical doctors and patients[15].However, ChatGPT revealed a 65% accuracy rate for both doctors and patients in the mentioned study. In our study, we did not apply any specific prompting due to the presence of three different models.

When evaluating the responses generated by LLMs, we recognized a totally incorrect answer for the question "Which antidepressants cause urinary incontinence?" This question was retrieved from www.answerthepublic.com, but in reality, there is no such relationship between antidepressants and urinary incontinence. Additionally, duloxetine, a serotonin-noradrenaline reuptake inhibitor, is currently being used by some regions of the world, including Europe, to treat stress urinary incontinence in women[16].

This type of inaccurate output generation, known as "hallucinations," is not rare among LLMs. It can have important consequences, especially in the context of health information for non-professionals[17].

LIMITATIONS OF THE STUDY

It's important to note a key limitation of our study, which is the subjective nature of the Likert scale used to assess the accuracy and comprehensiveness of the models' responses. Additonaly, LLMs are continuously evolving, so the results of the study should be considered within a specific timeline.

CONCLUSION

In conclusion, while all the tested LLMs are capable of generating health information related to urinary incontinence, our study reveals that GPT-4 demonstrates superior performance in both accuracy and comprehensiveness. However, it is crucial to exercise caution when using such models given the potential for generating incorrect information.

Conflicts of interest: The authors have nothing to disclose.

Funding: None.

REFERENCES

1. Abrams, P., Cardozo, L., Fall, M., Griffiths, D., Rosier, P., Ulmsten, U., Van Kerrebroeck, P., Victor, A., Wein, A. (2003). The standardisation of terminology in lower urinary tract function: report from the standardisation sub-committee of the International Continence Society. *Urology*, 61(1), 37–49.

2. Teunissen, D., Van Den Bosch, W., Van Weel, C., Lagro-Janssen, T. (2006). "It can always happen": The impact of urinary incontinence on elderly men and women. *Scandinavian Journal of Primary Health Care*, 24(3), 166–73.

3. Abrams, P., Andersson, K. E., Birder, L., Brubaker, L., Cardozo, L., Chapple, C., Cottenden, A., Davila, W., de Ridder, D., Dmochowski, R., Drake, M., Dubeau, C., Fry, C., Hanno, P., Smith, J. H., Herschorn, S., Hosker, G., Kelleher, C., Koelbl, H., … (2010). Fourth International Consultation on Incontinence Recommendations of the International Scientific Committee: Evaluation and treatment of urinary incontinence, pelvic organ prolapse, and fecal incontinence. *NeurourolUrodyn*, 29(1), 213–40.

4. Shaw, C., Tansey, R., Jackson, C., Hyde, C., Allan, R. (2001). Barriers to help seeking in people with urinary symptoms. *Family Practice*, 18(1), 48–52.

5. Moretti, FA., Oliveira, VE de., Silva, EMK da. (2012). Access to health information on the internet: a public health issue? *Rev Assoc Med Bras* (1992), 58(6), 650–8.

6. Wasserman, M., Baxter, NN., Rosen, B., Burnstein, M., Halverson, AL. (2014). Systematic Review of Internet Patient Information on Colorectal Cancer Surgery. *Diseases of the Colon & Rectum*, 57(1), 64–9.

7. Stacey, D., Légaré, F., Lewis, K., Barry, M. J., Bennett, C. L., Eden, K. B., Holmes-Rovner, M., Llewellyn-Thomas, H., Lyddiatt, A., Thomson, R., Trevena, L. (2017). Decision aids for people facing health treatment or screening decisions. *Cochrane Database of Systematic Reviews,* 2017(4).

8. Meskó, B., Topol, EJ. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *Npj Digit Med*, 6(1), 120.

9. Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*, 11(6), 887.

10. Elwyn, G., Frosch, D., Thomson, R., Joseph-Williams, N., Lloyd, A., Kinnersley, P., Cording, E., Tomson, D., Dodd, C., Rollnick, S., Edwards, A., Barry, M.(2012). Shared decision making: a model for clinical practice. *J Gen Intern Med*, 27(10), 1361–7.

11. Coskun, B., Ocakoglu, G., Yetemen, M., Kaygisiz, O. (2023). Can ChatGPT, an Artificial Intelligence Language Model, Provide Accurate and High-quality Patient Information on Prostate Cancer?. *Urology*, S0090-4295(23)00570-8. Advance online publication.

12. Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E., Scoville, E., Reese, T., Friedman, D., Bastarache, J., Heijden, Y. V. D., Wright, J., Carter, N., Alexander, M., Choe, J., … Wheless, L. (2023). Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. In Review. Available from: https://www.researchsquare.com/article/rs-2566942/v1

13. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D.(2020). Language Models are Few-Shot Learners*. arXiv.* Available from: http://arxiv.org/abs/2005.14165

14. Gabriel, J., Shafik, L., Alanbuki, A., Larner, T. (2023). The utility of the ChatGPT artificial intelligence tool for patient education and enquiry in robotic radical prostatectomy. *Int Urol Nephrol*. Available from: https://link.springer.com/10.1007/s11255-023-03729-4

15. Kaarre, J., Feldt, R., Keeling, LE., Dadoo, S., Zsidai, B., Hughes, JD., et al. (2023). Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. Knee Surg Sports TraumatolArthrosc. Available from: https://link.springer.com/10.1007/s00167-023-07529-2

16. Naumann, G., Aigmüller, T., Bader, W., Bauer, R., Beilecke, K., Betschart Meier, C., Bruer, G., Bschleipfer, T., Deniz, M., Fink, T., Gabriel, B., Gräble, R., Grothoff, M., Haverkamp, A., Hampel, C., Henscher, U., Hübner, M., Huemer, H., Kociszewski, J., … Reina, T. (2023). Diagnosis and Therapy of Female Urinary Incontinence. Guideline of the DGGG, OEGGG and SGGG (S2k-Level, AWMF Registry No. 015/091, January 2022): Part 1 with Recommendations on Diagnostics and Conservative and Medical Treatment. *GeburtshilfeFrauenheilkd*, 83(04), 377–409.

17. Coskun, B., Ocakoglu, G., Yetemen, M., Kaygisiz, O. (2023). AUTHOR REPLY. Urology, S0090-4295(23)00572-1.