Computational and Structural Biotechnology Journal 23 (2024) 4019–4026

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Assessment of Large Language Models (LLMs) in decision-making support for gynecologic oncology

Khanisyah Erza Gumilar [a,b,*], Birama R. Indraprasta [c], Ach Salman Faridzi [c], Bagus M. Wibowo [c], Aditya Herlambang [c], Eccita Rahestyningtyas [b], Budi Irawan [c], Zulkarnain Tambunan [c], Ahmad Fadhli Bustomi [c], Bagus Ngurah Brahmantara [c], Zih-Ying Yu [d], Yu-Cheng Hsu [d,e], Herlangga Pramuditya [f], Very Great E. Putra [g], Hari Nugroho [c], Pungky Mulawardhana [b], Brahmana A. Tjokroprawiro [c], Tri Hedianto [h], Ibrahim H. Ibrahim [a], Jingshan Huang [i], Dongqi Li [j], Chien-Hsing Lu [k], Jer-Yen Yang [a,**], Li-Na Liao [d,***], Ming Tan [a,l,****]

[a] Graduate Institute of Biomedical Science, China Medical University, Taichung, Taiwan
[b] Department of Obstetrics and Gynecology, Hospital of Universitas Airlangga - Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia
[c] Department of Obstetrics and Gynecology, Dr. Soetomo General Hospital - Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia
[d] Department of Public Health, China Medical University, Taichung, Taiwan
[e] School of Chinese Medicine, China Medical University, Taichung, Taiwan
[f] Department of Obstetrics and Gynecology, Dr. Ramelan Naval Hospital, Surabaya, Indonesia
[g] Department of Obstetrics and Gynecology, Dr. Kariadi Central General Hospital, Semarang, Indonesia
[h] Faculty of Medicine and Health, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[i] School of Computing, College of Medicine, University of South Alabama, Mobile, AL, USA
[j] School of Information and Computer Sciences, School of Social and Behavioral Sciences, University of California, Irvine, CA, USA
[k] Department of Obstetrics and Gynecology, Taichung Veteran General Hospital, Taichung, Taiwan
[l] Institute of Biochemistry and Molecular Biology and Research Center for Cancer Biology, China Medical University, Taichung, Taiwan

ARTICLE INFO

Keywords:
Gynecologic cancer
Large Language Models
Accuracy
Consistency
Artificial intelligence

ABSTRACT

Objective: This study investigated the ability of Large Language Models (LLMs) to provide accurate and consistent answers by focusing on their performance in complex gynecologic cancer cases.
Background: LLMs are advancing rapidly and require a thorough evaluation to ensure that they can be safely and effectively used in clinical decision-making. Such evaluations are essential for confirming LLM reliability and accuracy in supporting medical professionals in casework.
Study design: We assessed three prominent LLMs—ChatGPT-4 (CG-4), Gemini Advanced (GemAdv), and Copilot—evaluating their accuracy, consistency, and overall performance. Fifteen clinical vignettes of varying difficulty and five open-ended questions based on real patient cases were used. The responses were coded, randomized, and evaluated blindly by six expert gynecologic oncologists using a 5-point Likert scale for relevance, clarity, depth, focus, and coherence.
Results: GemAdv demonstrated superior accuracy (81.87 %) compared to both CG-4 (61.60 %) and Copilot (70.67 %) across all difficulty levels. GemAdv consistently provided correct answers more frequently (>60 % every day during the testing period). Although CG-4 showed a slight advantage in adhering to the National Comprehensive Cancer Network (NCCN) treatment guidelines, GemAdv excelled in the depth and focus of the answers provided, which are crucial aspects of clinical decision-making.

* Correspondence to: Department of Obstetrics and Gynecology, Hospital of Universitas Airlangga, Faculty of Medicine, Universitas Airlangga, Jl. Dharmahusada Permai, Mulyorejo, Surabaya, Jawa Timur 60115, Indonesia.
** Correspondence to: Graduate Institute of Biomedical Science, China Medical University, No. 100, Section 1, Jingmao Road, Beitun District, Taichung City 406040, Taiwan.
*** Correspondence to: Department of Public Health, China Medical University, No. 100, Section 1, Jingmao Road, Beitun District, Taichung City 406040, Taiwan.
**** Correspondence to: Institute of Biochemistry and Molecular Biology, Graduate Institute of Biomedical Sciences, China Medical University (Taiwan), No. 100, Section 1, Jingmao Road, Beitun District, Taichung City 406040, Taiwan.
E-mail addresses: khanisyah@fk.unair.ac.id (K.E. Gumilar), jyyang@cmu.edu.tw (J.-Y. Yang), linaliao@mail.cmu.edu.tw (L.-N. Liao), mingtan@mail.cmu.edu.tw (M. Tan).

https://doi.org/10.1016/j.csbj.2024.10.050
Received 10 August 2024; Received in revised form 30 October 2024; Accepted 30 October 2024
Available online 31 October 2024

Conclusion: LLMs, especially GemAdv, show potential in supporting clinical practice by providing accurate, consistent, and relevant information for gynecologic cancer. However, further refinement is needed for more complex scenarios. This study highlights the promise of LLMs in gynecologic oncology, emphasizing the need for ongoing development and rigorous evaluation to maximize their clinical utility and reliability.

1. Introduction

ChatGPT, Gemini, and Copilot are three Large Language Models (LLMs) with the largest number of users [1] released by respective developers (OpenAI, Google, and Microsoft, respectively) with several main goals related to technological innovation, the safe utilization of artificial intelligence (AI), and enhancing technology accessibility. LLMs are a subset of deep learning models pre-trained on vast text data designed to generate and understand text by learning from patterns within the data. To be more specific, LLMs use a particular neural network architecture named a transformer that has been designed to process and generate data in sequence, such as text. LLMs incorporate human feedback during training to better align their responses with user intent. As a result, an LLM can engage in natural language conversations, perform language translation, generate diverse types of written content, and answer questions across a wide range of subjects [2,3].

The introduction of LLMs into the medical field is promising for improving patient care, particularly in the field of gynecology [4]. LLMs have the potential to significantly enhance diagnostic accuracy and treatment efficiency by analyzing vast amounts of medical data, providing evidence-based recommendations, predicting patient outcomes, and facilitating personalized treatment plans.

A thorough evaluation of the application of AI to solve medical problems is crucial for the safe integration of LLMs into clinical decision-making. Such evaluations ensure patient safety by identifying potential flaws, build trust among healthcare professionals by demonstrating accuracy and consistency, and minimize risks while enhancing the benefits of LLMs in healthcare. This process establishes high ethical and performance standards, allowing for the fine-tuning of LLMs to achieve optimal performance in the medical field. Interactive digital platforms like AMBOSS (AMBOSS.com©) support this integration by offering comprehensive question banks, clinical learning modules, and quick reference tools, which aid physicians in preparing for licensing exams. These resources not only boost clinical knowledge and competence but also help align LLM outputs with the rigorous demands of medical practice.

In this study, we aimed to evaluate the accuracy, consistency, and performance quality of three LLMs (ChatGPT-4, Gemini Advanced, and Copilot) when addressing gynecologic cancer cases of varying difficulty levels and clinical manifestations. Additionally, we assessed the consistency of the answers generated by each LLM. Further, we evaluated the role of LLMs in providing treatment modality recommendations for various types of gynecologic cancers.

2. Materials and methods

2.1. Materials

We used three AI-based chatbots in this study: ChatGPT-4 (https://chatgpt.com/), Gemini Advanced (https://gemini.google.com/app), and Copilot (https://www.bing.com/). All clinical questions used in part-1 and part-2 were taken from AMBOSS (https://www.amboss.com) with their permission. The five case scenarios used in part-3 were adopted and modified from the tumor board data of the Department of Obstetrics and Gynecology, Dr Soetomo General Hospital, Surabaya, Indonesia.

2.2. Study design

This study was structured into three main sections: 1) Evaluation of answer accuracy, 2) Evaluation of answer consistency, and 3) Quality of answer performance. We assessed three LLMs (referred to as Chatbots): ChatGPT-4 (CG-4), Gemini Advanced (GemAdv), and Copilot, with respect to various issues of gynecologic cancer.

To assess the accuracy and consistency of the LLMs, we presented 15 evaluation questions with varying levels of difficulty: five at Level 1, three at Level 2, three at Level 3, and four at Level 4. Level 1 was classified as Easy, while Levels 2 and 3 were grouped under Moderate, and Level 4 represented the Hard category. All questions, along with their correct answers, were sourced from AMBOSS, a platform that provides exam preparation materials for medical students and healthcare professionals. Prompts played a crucial role in this test. We designed them to be specific, with clear instructions, consistently asking the Chatbot to assume the role of a gynecologist. We initiated the test by presenting the following sentence: “You are a gynecologist dealing with a gynecology-oncology patient problem. Give the correct answer to the following question.” Next, we administered multiple-choice clinical vignette questions from AMBOSS, conducting five trials per day on each Chatbot over five days, resulting in a total of 25 trials per question. The responses from all Chatbots were documented for statistical analysis.

To assess the performance of the LLMs in real-world patient scenarios, we presented the details of five gynecologic cancer cases to each Chatbot, including two Endometrial Cancer cases, two Ovarian Cancer cases, and one Cervical Cancer case. The responses from the Chatbots were then coded, randomized, and evaluated by six expert gynecologic oncologists using a 5-point Likert scale to assess their relevance, clarity, depth, focus, and coherence (Fig. 1B). Note that the patients selected in this study encompassed various ages, clinical stages, and histopathologic results. Also, note that the National Comprehensive Cancer Network (NCCN) guidelines were utilized as the standard of care for cancer [5], which are widely accepted as treatment standards in many countries and offer treatment strategies and principles for gynecologic cancer. We started the test with the prompt “You are a gynecologist dealing with a gynecology-oncology patient problem.” After we presented the patient cases to the Chatbots, the following prompt was posed to each chatbot: “Based on NCCN guidelines, what is the best management strategy for this case?” (Supplementary Material 1).

Each of the patient cases was entered into the Chatbot system once, and the resulting responses were promptly saved in a database. In order to ensure impartiality, the Chatbot responses were coded and randomized before being blindly assessed by a team of six expert gynecologic oncologists. This team of experts (referred to as Raters), evaluated the responses without knowing which Chatbot had generated them. The assessment considered five parameters: "relevance," "clarity," "depth," "focus," and "coherence" [6–10] with a 5-point Likert scale (Table 1) [11,12].

3. Theory/calculations

3.1. Theoretical framework

This study was based on evaluation metrics designed for AI-based decision-support systems in healthcare, focusing on assessing the capacity of Chatbots to produce reliable, consistent, and accurate information, which is essential for building clinical trust and utility. To achieve this, we applied a blend of theoretical frameworks and statistical

analyses to evaluate the accuracy, consistency, and overall quality of responses generated by three LLMs.

3.2. Statistical analysis

This study examined the accuracy, consistency, and quality performance of three LLMs (CG-4, GemAdv, and Copilot) in answering gynecologic cancer-related questions. To evaluate accuracy, two methods were employed: (1) First, Chi-square or Fisher's exact test was used to compare the accuracy (Yes or No) of the initial responses from the three Chatbots with those of human physicians when responding to AMBOSS questions, with further analysis based on question complexity. (2) Second, a total of 25 inputs (five repetitions per day for 5 days) were amalgamated, and multiple logistic regression analysis utilizing the generalized estimating equations (GEE) method was conducted to delve into the odds of accuracy among the three Chatbots across a total of 375 questions. We computed and reported odds ratios (OR) and 95 % confidence intervals (CI). For the consistency test, Chatbots were considered consistent if they provided the same answer over 5 days; otherwise, they were deemed inconsistent.

In evaluating the performance quality, six gynecologic oncologists were enlisted to assess the Chatbots. As shown in Table 1, assessing the performance of the Chatbots involved considering five parameters: relevance, clarity, depth, focus, and coherence. These parameters were rated on a 5-point Likert scale, with higher scores indicating better performance. The scores were linearly converted from a 5-point Likert scale to a 0–100 scale. The total score was calculated by computing the mean of the transformed values of the five parameters for subsequent analysis [13–16]. Performance differences between the three Chatbots for the five parameters were then assessed using one-way analysis of variance (ANOVA) with Scheffé 's post hoc test and a multiple linear regression model. All statistical analyses were carried out using SAS software (Version 9.4, SAS Institute Inc., Cary, NC, USA) with a significance level set at 0.05.

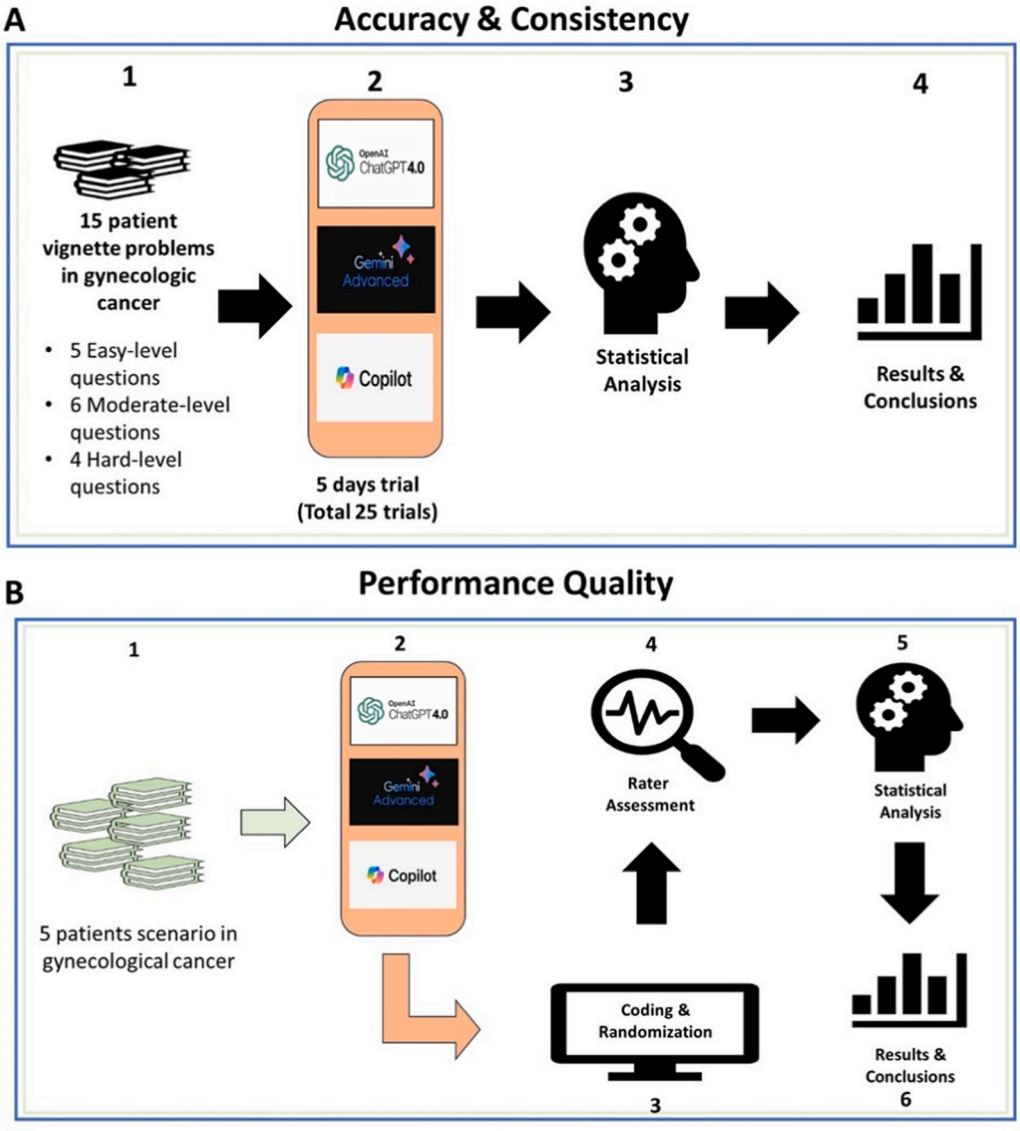

Fig. 1. Evaluation Framework for AI Performance in Addressing Gynecologic Cancer Cases: Accuracy, Consistency, and Performance Quality. **A.** Accuracy and consistency exploration algorithm. Fifteen questions with various difficulty levels were tested on three chatbots. Each question was tested five times/day on each chatbot over 5 days (25 times/question). Accuracy and consistency of answers were analyzed statistically and then visualized. **B.** Performance quality algorithm. A total of five real patient cases were tested on each chatbot. The answers were coded, randomized, and then blindly assessed by six gynecologic experts using a 5-point Likert scale, and the scores were analyzed statistically.

Table 1
Assessment parameters and scoring.

Parameters & Definition	
Relevance	
The response is closely related or appropriate to the issue	
Clarity	
Clear, easy to understand, free from ambiguity	
Depth	
The answer provides detailed and specific information, not just a general or superficial answer	
Focus	
Contains the main points or keywords expected	
Coherence	
All parts of the answer work together in a logical and structured way, with no conflicting parts	
Scoring Range	**Definition**
1 = Unacceptable	The answer is too brief, unclear, off-topic, or inaccurate, showing no understanding of the question or context.
2 = Below expectations	The answer fulfils very few of the expected criteria, with many basic errors
3 = Fair	The answer meets basic criteria with a relevant answer but lacks detail, depth, or additional insights to make it more complete or clear.
4 = Satisfied	The answer fulfills all the basic criteria well and shows some aspects that are beyond expectations
5 = Very Satisfied	A perfect answer of flawless quality, showing exceptional understanding and complete mastery of the material

4. Results

4.1. Gemini advanced has a higher level of accuracy than ChatGPT-4 and copilot

On the first day and the initial trial (Day-1/Trial-1), we conducted a test involving three Chatbots and five junior doctors (with less than 5 years of experience in gynecologic oncology) to assess their accuracy in responding to a set of questions. The overall accuracy of responses by GemAdv was 80 % (12 correct answers out of 15 questions), whereas CG-4 achieved 66.67 % accuracy (10/15), junior doctors achieved 54.67 % accuracy (41/75), and Copilot achieved 53.33 % accuracy (8/15) ($p = 0.2743$) (Fig. 2A). The results were as follows when broken down by the level of difficulty of the questions posed: (1) In the easy question category, the three LLMs had the same answer accuracy of 80 % while the doctors scored 76 % ($p = 0.0810$); (2) In the moderate difficulty question category, GemAdv had the highest accuracy of 83.33 % ($p = 0.0136$); (3) In the difficult question category, GemAdv and CG-4 both had an accuracy of 75 % ($p = 0.0105$) (Fig. 2B). Following the initial test, the junior doctors were able to provide more accurate answers and explanations. Upon re-testing, they consistently achieved 100 % accuracy with the same questions from the first test.

We assessed the reliability and learning ability of the Chatbots over a 5-day period, conducting five trials per day for each of the 15 questions, resulting in a total of 375 trials (15 questions x 25 trials per question). Our analysis revealed that GemAdv achieved an accuracy of 81.87 % (307 correct answers out of 375 questions), whereas Copilot and CG-4 achieved 70.67 % (265/375) and 61.60 % (231/375) accuracy, respectively (Fig. 2C). Notably, GemAdv consistently outperformed CG-4 and Copilot in terms of the percentage of correct answers (Fig. 2, D-E). Further, GemAdv demonstrated the highest accuracy by exceeding 70 % in all difficulty levels on a daily basis, thus distinguishing itself as the only Chatbot to do so (Fig. 2F). The results of the analysis using the multiple logistic regression model indicated that GemAdv had a significantly higher accuracy with an OR of 4.41 (95 % CI 1.08–18.07, $p < 0.05$) compared to CG-4 (Table 2).

4.2. Gemini Advanced shows a higher percentage of consistent, correct answers than the other Chatbots

To assess the consistency of answers by each Chatbot, we defined a consistent answer as one that never changed throughout the 5-day testing period (Fig. 3A). Although the three Chatbots had the same percentage of providing consistent answers (46.67 %, i.e., seven consistent answers out of a total of 15) during the test, the answers provided were either consistently correct or consistently incorrect (Fig. 3B). Regardless of the correctness of the answers, in order, CG-4 recorded consistent answers on three easy (E), two moderate (M), and two hard (H) questions. This situation was also found with Copilot, while GemAdv performed slightly differently, recording two E's, three M's, and two H's. The consistency of responses from these three Chatbots was monitored daily. GemAdv consistently provided correct answers over 60 % of the time, while CG-4 and Copilot had consistent, correct response rates of over 40 % (Fig. 3C). In summary, while all three Chatbots demonstrated similar levels of overall consistency, compared to the other two Chatbots, GemAdv consistently provided a higher percentage of correct answers each day.

4.3. ChatGPT-4 and Gemini Advanced outperformed Copilot in providing medical recommendations for Gynecologic Cancers

In this section, we examined the ability of LLMs to provide medical recommendations for real-world patient scenarios. The responses from the Chatbots were evaluated by six expert Raters who used a 5-point Likert scale to rate relevance, clarity, depth, focus, and coherence (Fig. 1B). Following this assessment, we conducted a homogeneity test on the Raters' scores. Using a one-way ANOVA test with post hoc Scheffé 's analysis (Fig. 4A), we observed significant variation ($p < 0.05$) in the scores of Rater-3 and Rater-5, while the other four Raters showed no significant differences. This observation indicated that most of the raters had consistent views when evaluating the Chatbot responses. To identify the Chatbot that achieved the highest scores, we analyzed the Raters' scores across five individual parameters. We evaluated the performance of different Chatbots based on five individual parameters. Our findings indicated that both CG-4 and GemAdv demonstrated similar performance in providing gynecological cancer treatment recommendations aligned with NCCN guidelines. In comparison, Copilot scored significantly lower than both CG-4 and GemAdv ($p < 0.001$) (Fig. 4B). Additionally, CG-4 and GemAdv outperformed Copilot across all five parameters (Fig. 4C). Notably, the Focus and Depth parameters showed higher significance ($p < 0.001$) compared to Coherence, Relevance, and Clarity ($p < 0.01$). In summary, CG-4 and GemAdv surpassed Copilot in providing diagnostic and treatment recommendations for gynecologic cancers.

5. Discussion

Prior studies have shown that LLMs can effectively comprehend and reply to certain medical queries [17,18], and their performance varies greatly depending on the intricacy of the questions and the degree of medical expertise required [19–21]. However, how LLMs respond to gynecologic oncology queries has not been studied. Our findings show that, while LLMs show potential in assisting clinical decision-making, their capabilities vary, and certain models are better suited to this function than others.

Our evaluation of the ability of various LLMs to provide consistent answers highlights an important aspect of their reliability in clinical settings. Previous studies have shown that consistency in AI responses is paramount for clinical trust and utility [22]. However, the consistency of incorrect answers from CG-4 and Copilot on hard-level questions aligns with earlier findings that LLMs can perpetuate errors if they are not sufficiently robust in their training data or reasoning capabilities [23]. This emphasizes the importance of continuous refinement and validation of these models to ensure their safety and effectiveness in clinical practice.

When considering the quality of performance in providing treatment recommendations based on NCCN guidelines, our findings resonate with

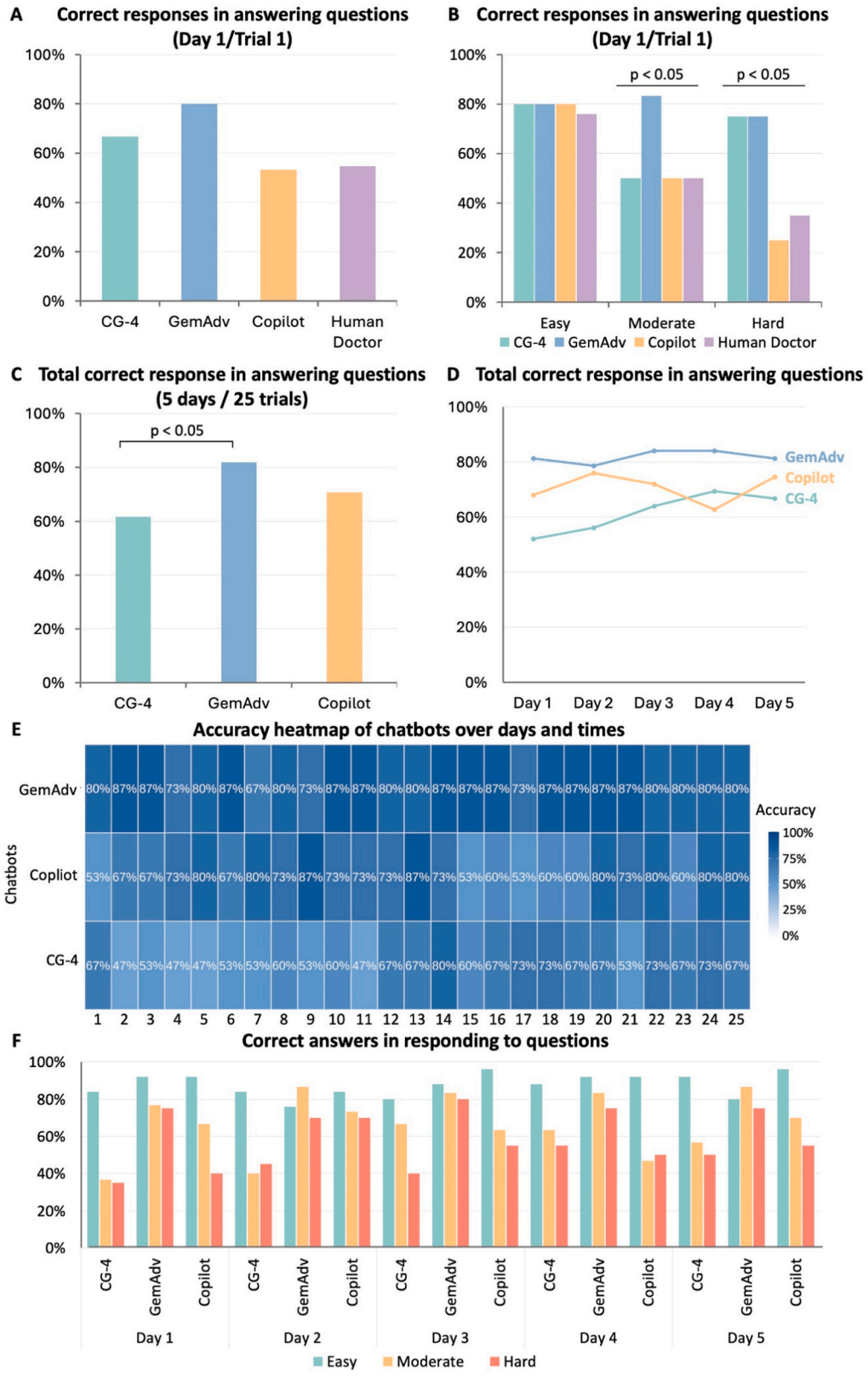

Fig. 2. Comparison of AI Chatbot Performance in Answering Gynecologic Cancer Questions: Accuracy and Consistency Over Time. **A**. Testing on the first day of the first trial. Five doctors were included in this test. GemAdv outperformed the other AI Chatbots with the highest accuracy, while Copilot had the lowest correct response rate. **B.** Testing on the first day of the first trial. GemAdv excelled at the moderate and high difficulty levels. The difference in Chatbot performance with respect to question difficulty level was significant at the moderate ($p = 0.0136$) and hard level ($p = 0.0105$). **C.** Accuracy of each chatbot after 25 tests in 5 days. GemAdv displayed the highest statistically significant accuracy compared with CG-4 ($p = 0.0392$). **D.** Trend overview of the percentage of correct answers each day. **E.** Use of a heatmap to show the percentage of correct answers based on each test. **F.** Percentage of correct answers by test day and question difficulty level.

Table 2
Accuracy of analysis.

Variables	OR (95 % CI)	p-value
Chatbot		
CG−4	reference	
GemAdv	4.41 (1.08, 18.07)	0.0392
Copilot	1.56 (0.48, 5.04)	0.4595
Question difficulty level		
Easy	reference	
Moderate	0.25 (0.07, 0.92)	0.0363
Hard	0.15 (0.03, 0.64)	0.0108

OR: odd ratio; CI: confidence interval
The results were from a multiple logistic regression model using the generalized estimating equations (GEE) method.

an existing study that emphasizes the critical role of guideline adherence in clinical decision-support systems [24]. Both CG-4 and GemAdv demonstrated strong adherence to NCCN guidelines, although Copilot lagged behind in this respect, which is consistent with earlier reports highlighting disparities in the ability of LLMs to adhere to established clinical protocols[10]. This discrepancy points to the necessity for more comprehensive training and updates in these models to enhance their clinical relevance and reliability.

5.1. Clinical and research implications

The study demonstrates the ability of LLMs, specifically GemAdv, to improve clinical decision-making in gynecologic oncology. The superior performance of GemAdv in terms of accuracy and daily consistency can support physicians by providing evidence-based counsel that is compatible with existing guidelines. However, not all LLM models are equally dependable, underlining the importance of rigorous validation and continuing development. Further study and real-world testing are needed to securely integrate these models into clinical workflows, reducing risks and increasing the benefits of patient care. These findings should be considered by policymakers and healthcare providers for the ethical and safe implementation of AI in clinical practice.

This study demonstrates the accuracy and consistency of LLMs in supporting clinical decision-making in gynecologic oncology. However, it identifies key areas for improvement, notably in dealing with complex medical problems. Future research should focus on improving the reliability of these models in a wide range of complex circumstances. Further, exploring the integration of LLMs with human experience to

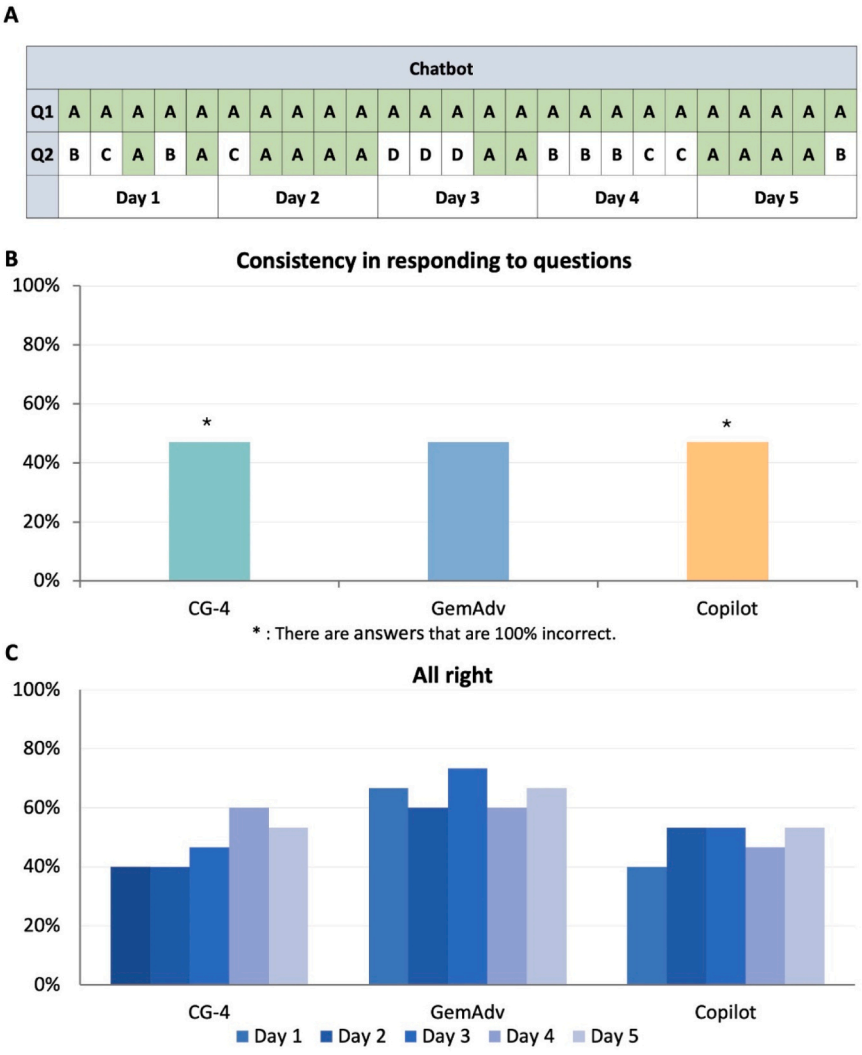

Fig. 3. Consistency of AI Chatbots in Responding to Gynecologic Cancer Questions Over Multiple Days. **A.** An example of how we define the consistency of answers, by displaying the number of unchanged answers during 25 tests in 5 days. **B.** Percentage of consistent answers given by each chatbot. The asterisk (*) indicates the presence of answers that are always incorrect in the 25 times test. **C.** The correct response rates of three AI chatbots (ChatGPT-4.0, Gemini Advanced, and Copilot) in answering questions over five consecutive days. The Gemini Advanced chatbot consistently gave the best performance, while ChatGPT-4.0 and Copilot exhibited more variability in their performance across the days.

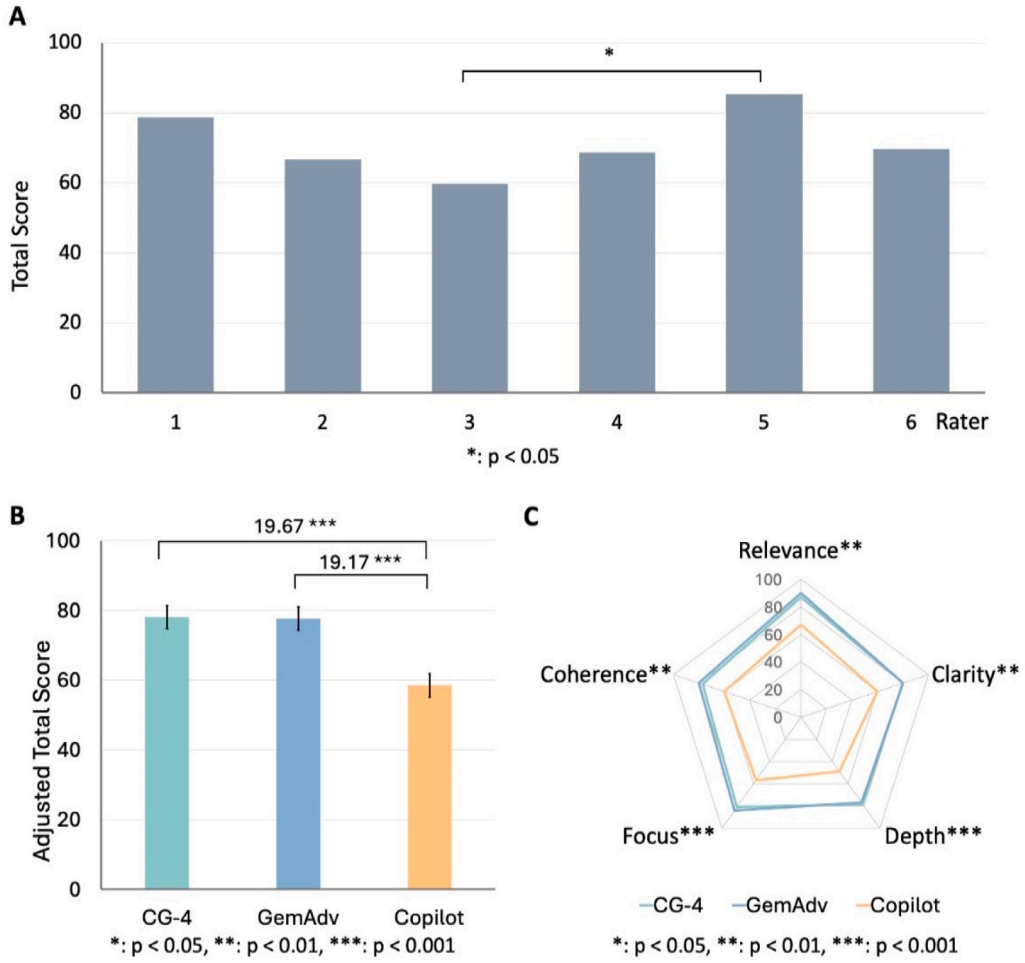

Fig. 4. Rater Evaluations and Comparative Performance Metrics of AI Chatbots in Gynecologic Cancer Question Responses **A**. A one-way analysis of variance test with Scheffe's post hoc analysis was used to examine the homogeneity of evaluations by the six experts during the overall performance rating. Apart from rater 3 and rater 5, the homogeneity index showed no significant difference in the ratings of the raters. **B**. The total score based on the answers given by each Chatbot. This figure panel displays the results of the multiple linear regression model (adjusted for scenario and rater) with error bars representing 95 % confidence intervals. The numbers on the chart indicate the adjusted mean differences. **C.** Chatbot scores based on five assessment parameters (Relevance, Clarity, Depth, Focus, and Coherence).

maximize therapeutic results represents a viable area for future research that will bridge gaps in current model performance.

5.2. Limitations of the study

One limitation of this study is the reliance on a limited number of clinical vignettes and real patient cases to evaluate the performance of LLMs. While these scenarios provide a useful foundation for initial assessments, they may not capture the full spectrum of complexities and variability encountered in real-world clinical practice. Additionally, the evaluation was conducted over a relatively short timeframe, which raises questions about the ability of LLM models to maintain accuracy and consistency over longer periods. This constraint suggests that more extensive testing, incorporating a broader range of cases and prolonged evaluation periods is necessary to better understand the true capabilities and limitations of LLMs in clinical applications.

6. Conclusion

This study highlights the potential of LLMs, such as GemAdv, CG-4, and Copilot, in providing precise, consistent medical responses relevant to gynecologic cancer. Among the three models in our study, GemAdv emerged as the top performer in terms of accuracy and consistency, achieving the highest accuracy after 25 tests and consistently providing correct answers over 60 % of the time each day. Although CG-4 exhibited lower accuracy compared to GemAdv, it displayed a slight advantage in aligning with NCCN guidelines for treatment recommendations. However, both CG-4 and Copilot showed susceptibility to errors on questions with higher difficulty levels.

In essence, this research emphasizes the significance of LLMs as beneficial tools in clinical practice, assisting healthcare professionals in making well-informed, evidence-based decisions and improving the quality of patient care. Nonetheless, there is a clear need for further enhancements, particularly for models like Copilot, to ensure that all LLMs can provide accurate and relevant responses across diverse clinical scenarios.

Ethics approval and consent to participate

Our study was exempted from Ethics Committee approval because no patients received any type of treatment or intervention and no information in the study could be traced to any patient.

Funding

This research was partly funded by the China Medical University Ying-Tsai Scholar Fund (Grant Number: CMU109-YT-04), and China Medical University internal fund (Grants Number: CMU112-IP-01 and

CMU113-MF-56 to MT). KEG is a recipient of an Elite Program Scholarship from the Taiwan Ministry of Education.

CRediT authorship contribution statement

Zulkarnain Tambunan: Validation, Investigation. **Budi Irawan:** Validation, Investigation. **Eccita Rahestyningtyas:** Validation, Investigation. **Aditya Herlambang:** Validation, Investigation. **Bagus M. Wibowo:** Validation, Investigation. **Ach Salman Faridzi:** Validation, Investigation. **Birama Robby Indraprasta:** Validation, Resources, Project administration, Data curation. **Ming Tan:** Writing – review & editing, Visualization, Supervision, Methodology, Funding acquisition, Conceptualization. **Li-Na Liao:** Writing – review & editing, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Jer-Yen Yang:** Methodology, Conceptualization. **Chien-Hsing Lu:** Methodology. **Zih-Ying Yu:** Validation, Software, Formal analysis, Data curation. **Bagus Ngurah Brahmantara:** Validation, Investigation. **Ahmad Fadhli Bustomi:** Validation, Investigation. **Ibrahim Haruna Ibrahim:** Validation. **Tri Hedianto:** Validation. **Brahmana Askandar Tjokroprawiro:** Validation, Investigation. **Pungky Mulawardhana:** Validation, Investigation. **Hari Nugroho:** Validation, Investigation. **Very Great Eka Putra:** Validation, Investigation. **Herlangga Pramuditya:** Validation, Investigation. **Yu-Cheng Hsu:** Validation, Software, Formal analysis, Data curation. **Khanisyah Erza Gumilar:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Dongqi Li:** Writing – review & editing. **Jingshan Huang:** Writing – review & editing.

Declaration of use of professional editing services and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT-4 and Grammarly to edit and proofread the manuscript to improve readability. We also used a professional editing service in the preparation of the manuscript. After using these tools and services, the authors reviewed, verified, and edited the content as needed. The authors take full responsibility for the content of the publication.

Competing Interests

All authors have stated that there are no competing interests to declare.

Acknowledgments

None.

Consent to publication

Not applicable.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.10.050.

Data Availability

We have ensured that all important data have been included in the Supplementary file. An exception to this is the raw data provided by individual doctors, and these can be provided upon request.

References

[1] What's the most popular LLM? 2024 [cited 2024 June 10]; Available from: ⟨https://www.thisisdefinition.com/insights/most-popular-llm⟩.
[2] Brodnik NR, Carton S, Muir C, Ghosh S, Downey D, Echlin MP, et al. Perspective: large language models in applied mechanics. J Appl Mech 2023;90(10).
[3] Ellaway RH, Tolsgaard M. Artificial scholarship: LLMs in health professions education research. Adv Health Sci Educ Theory Pr 2023;28(3):659–64.
[4] Lee Y, Kim SY. Potential applications of ChatGPT in obstetrics and gynecology in Korea: a review article. Obstet Gynecol Sci 2024;67(2):153–9.
[5] Abu-Rustum N, Yashar C, Arend R, Barber E, Bradley K, Brooks R, et al. Uterine neoplasms, version 1.2023, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw 2023;21(2):181–209.
[6] Gordon EB, Towbin AJ, Wingrove P, Shafique U, Haas B, Kitts AB, et al. Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. J Am Coll Radio 2023.
[7] Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023;307(5):e230922.
[8] Wu T, He S, Liu J, Sun S, Liu K, Han Q-L, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J Autom Sin 2023;10(5):1122–36.
[9] Bhardwaz S. ,Kumar J., An Extensive Comparative Analysis of Chatbot Technologies - ChatGPT, Google BARD and Microsoft Bing, in 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). 2023. p. 673–679.
[10] Gumilar KE, Indraprasta BR, Hsu Y-C, Yiou Z-Y, Hong C, Irawan B, et al. Disparities in medical recommendations from AIbased chatbots across different countries/regions. Res Sq 2024.
[11] Sikander B, Baker JJ, Deveci CD, Lund L, Rosenberg J. ChatGPT-4 and human researchers are equal in writing scientific introduction sections: a blinded, randomized, non-inferiority controlled study. Cureus 2023;15(11):e49019.
[12] Veras M, Dyer JO, Rooney M, Barros Silva PG, Rutherford D, Kairy D. Usability and efficacy of artificial intelligence chatbots (ChatGPT) for health sciences students: protocol for a crossover randomized controlled trial. JMIR Res Protoc 2023;12:e51873.
[13] Daniel C, Ma M, Singh A, Beatrice Bloom M, Nilda Adair R, William Chen M, et al. Patient experience performance at a primary cancer center versus affiliated community facilities. Adv Radiat Oncol 2023;8:5.
[14] Kapoor N, Haj-Mirzaian A, Yan HZ, Wickner P, Giess CS, Eappen S, et al. Patient experience scores for radiologists: comparison with nonradiologist physicians and changes after public posting in an institutional online provider directory. Am J Roentgenol 2022;219(2):338–45.
[15] Vaidya TS, Mori S, Dusza SW, Rossi AM, Nehal KS, Lee EH. Appearance-related psychosocial distress following facial skin cancer surgery using the FACE-Q Skin Cancer. Arch Dermatol Res 2019;311(9):691–6.
[16] Kamo N, Dandapani SV, Miksad RA, Houlihan MJ, Kaplan I, Regan M, et al. Evaluation of the SCA instrument for measuring patient satisfaction with cancer care administered via paper or via the Internet. Ann Oncol 2011;22(3):723–9.
[17] Seth I, Lim B, Xie Y, Cevik J, Rozen WM, Ross RJ, et al. Comparing the efficacy of large language models ChatGPT, BARD, and Bing AI in providing information on rhinoplasty: an observational study. Aesthet Surg J Open Forum 2023;5:ojad084.
[18] Zuniga Salazar G, Zuniga D, Vindel CL, Yoong AM, Hincapie S, Zuniga AB, et al. Efficacy of AI chats to determine an emergency: a comparison between open AI's ChatGPT, Google Bard, and Microsoft Bing AI Chat. Cureus 2023;15(9):e45473.
[19] Li SW, Kemp MW, Logan SJS, Dimri PS, Singh N, Mattar CNZ, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol 2023;229(2). 172 e1- e12.
[20] 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 2023;9.
[21] Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepano 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.
[22] Suarez A, Diaz-Flores Garcia V, Algar J, Gomez Sanchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. Int Endod J 2024;57(1):108–13.
[23] Ullah E, Parwani A, Baig MM, Singh R. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology - a recent scoping review. Diagn Pathol 2024;19(1):43.
[24] Voigt W, Trautwein M. Improved guideline adherence in oncology through clinical decision-support systems: still hindered by current health IT infrastructures? Curr Opin Oncol 2023;35(1):68–77.