Assess the capabilities of AI-based large language models (AI-LLMs) in interpreting histopathological slides and scientific figures: performance evaluation study

Khanisyah Erza Gumilar, Grace Ariani, Priangga Adi Wiratama, Rimbun Rimbun, Tri Hartini Yuliawati, Hong Chen, Ibrahim Haruna Ibrahim, Cheng-Han Lin, Tai-Yu Hung, Dewanti Anggrahini, Arya Satya Rajanagara, Zih-Ying Yu, Yu-Cheng Hsu, Erry Gumilar Dachlan, Jer-Yen Yang, Li-Na Liao, Ming Tan

Submitted to: Journal of Medical Internet Research
on: October 07, 2024

Table of Contents

Assess the capabilities of AI-based large language models (AI-LLMs) in interpreting histopathological slides and scientific figures: performance evaluation study

Khanisyah Erza Gumilar[1] MD; Grace Ariani[2] MD; Priangga Adi Wiratama[2] MD; Rimbun Rimbun[3] MD; Tri Hartini Yuliawati[3] MD, PhD; Hong Chen[4] MSc; Ibrahim Haruna Ibrahim[4] MSc; Cheng-Han Lin[4] BSc; Tai-Yu Hung[4]; Dewanti Anggrahini[5] MEng; Arya Satya Rajanagara[3] MD; Zih-Ying Yu[6]; Yu-Cheng Hsu[6]; Erry Gumilar Dachlan[7]; Jer-Yen Yang[4] PhD; Li-Na Liao[8] PhD; Ming Tan[9] Prof Dr Med, PhD

[1]Department of Obstetrics and Gynecology Hospital of Universitas Airlangga - Faculty of Medicine Universitas Airlangga Surabaya ID
[2]Department of Pathology Anatomy Faculty of Medicine Universitas Airlangga Surabaya ID
[3]Department of Anatomy, Histology, and Pharmacology Faculty of Medicine Universitas Airlangga Surabaya ID
[4]Graduate Institute of Biomedical Science China Medical University Taichung City TW
[5]Department of Industrial & Systems Engineering Institut Teknologi Sepuluh Nopember Surabaya ID
[6]Department of Public Health College of Public Health China Medical University Taichung City TW
[7]Department of Obstetrics and Gynecology Faculty of Medicine Universitas Airlangga Surabaya ID
[8]Department of Public Health China Medical University Taichung City TW
[9]Institute of Biochemistry and Molecular Biology Graduate Institute of Biomedical Sciences China Medical University Taichung City TW

Corresponding Author:
Ming Tan Prof Dr Med, PhD
Institute of Biochemistry and Molecular Biology
Graduate Institute of Biomedical Sciences
China Medical University
No. 100, Sec. 1, Jingmao Rd, Beitun Dist
Taichung City
TW

Abstract

Background: Interpreting histopathology slides and scientific figures requires specialized skills and knowledge. Pathologists analyze various tissues and cells, while the general population often struggles with the technical information in scientific figures. Artificial intelligence-based large language models (AI-LLMs) can simplify these processes by providing clearer explanations.

Objective: This study explores the capabilities AI-LLMs in interpreting histopathology slides and scientific figures. The objective is to assess the value of AI LLMs in medical applications and scientific education.

Methods: The study was divided into two parts: interpreting histopathology slides and scientific figures. Six histopathology images and six scientific figures were tested on each of the three most frequently used chatbots (ChatGPT-4, Gemini Advanced, and Copilot). Responses from the chatbots were coded and blindly examined by expert raters using five parameters—relevance, clarity, depth, focus, and coherence—on a 5-point Likert scale. Statistical analysis included one-way ANOVA and multiple linear regression.

Results: ChatGPT-4 outperformed Gemini Adv and Copilot in both histopathology and scientific image interpretation, with significantly higher scores across all parameters (P<.001). High homogeneity among raters validated these findings. ChatGPT-4's superior performance may be due to its advanced algorithms, extensive training data, specialized modules, and user feedback.

Conclusions: ChatGPT-4 excels in interpreting histopathology and scientific images, which may lead to improving diagnostic accuracy, clinical decision-making, and reducing pathologists' workload. It also benefits education by enhancing students' understanding of complex images and promoting interactive learning. ChatGPT-4 shows a significant potential to improve patient care and enrich student learning.

(JMIR Preprints 07/10/2024:67270)
DOI: https://doi.org/10.2196/preprints.67270

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Assess the capabilities of AI-based large language models (AI-LLMs) in interpreting histopathological slides and scientific figures: performance evaluation study

Khanisyah E. Gumilar[1,2,*], Grace Ariani[3], Priangga A. Wiratama[3], Rimbun[4], Tri H. Yuliawati[4], Hong Chen [1], Ibrahim H. Ibrahim[1], Chen-Hang Lin[1], Tai-Yu Hung[1], Dewanti Anggrahini[5,6], Arya S. Rajanagara[1,4], Zih-Ying Yu[7], Yu-Cheng Hsu[7,8], Erry G. Dachlan[9], Jer-Yen Yang[1,10], Li-Na Liao[8*], Ming Tan[1,10,*]

1. Graduate Institute of Biomedical Science, China Medical University, Taichung, Taiwan, R.O.C
2. Department of Obstetrics and Gynecology, Universitas Airlangga Hospital- Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia
3. Department of Pathology Anatomy, Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia
4. Department of Anatomy, Histology and Pharmacology, Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia
5. Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu, Taiwan, R.O.C
6. Department of Industrial & Systems Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
7. Department of Public Health, China Medical University, Taichung, Taiwan, R.O.C
8. School of Chinese Medicine, China Medical University, Taichung, Taiwan, R.O.C
9. Department of Obstetrics and Gynecology, Dr Soetomo General Hospital - Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia
10. Institute of Biochemistry and Molecular Biology and Research Center for Cancer Biology, China Medical University, Taichung, Taiwan, R.O.C

*Corresponding author
Khanisyah Erza Gumilar
Department of Obstetrics and Gynecology, Hospital of Universitas Airlangga
Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia
Jl. Dharmahusada Permai, Mulyorejo, Kec. Mulyorejo, Surabaya, Jawa Timur 60115
Email: khanisyah@fk.unair.ac.id

Li-Na Liao
Department of Public Health, China Medical University, Taichung, Taiwan
No. 100, Sec. 1, Jingmao Rd, Beitun Dist,Taichung City 406040, Taiwan R.O.C.
Email: linaliao@mail.cmu.edu.tw

Ming Tan
Institute of Biochemistry and Molecular Biology, Graduate Institute of Biomedical Sciences, China Medical University (Taiwan)
No. 100, Sec. 1, Jingmao Rd, Beitun Dist,Taichung City 406040, Taiwan R.O.C.
Email: mingtan@mail.cmu.edu.tw

Abstract

Background: Interpreting histopathology slides and scientific figures requires specialized skills and knowledge. Pathologists analyze various tissues and cells, while the general population often

struggles with the technical information in scientific figures. Artificial intelligence-based large language models (AI-LLMs) can simplify these processes by providing clearer explanations.

Objectives: This study explores the capabilities AI-LLMs in interpreting histopathology slides and scientific figures. The objective is to assess the value of AI LLMs in medical applications and scientific education.

Methods: The study was divided into two parts: interpreting histopathology slides and scientific figures. Six histopathology images and six scientific figures were tested on each of the three most frequently used chatbots (ChatGPT-4, Gemini Advanced, and Copilot). Responses from the chatbots were coded and blindly examined by expert raters using five parameters—relevance, clarity, depth, focus, and coherence—on a 5-point Likert scale. Statistical analysis included one-way ANOVA and multiple linear regression.

Results: ChatGPT-4 outperformed Gemini Adv and Copilot in both histopathology and scientific image interpretation, with significantly higher scores across all parameters (P<.001). High homogeneity among raters validated these findings. ChatGPT-4's superior performance may be due to its advanced algorithms, extensive training data, specialized modules, and user feedback.

Conclusions: ChatGPT-4 excels in interpreting histopathology and scientific images, which may lead to improving diagnostic accuracy, clinical decision-making, and reducing pathologists' workload. It also benefits education by enhancing students' understanding of complex images and promoting interactive learning. ChatGPT-4 shows a significant potential to improve patient care and enrich student learning.

Keywords: AI-LLM, Artificial Intelligence, Large language model, Histopathological Image, Scientific Figure, ChatGPT, Gemini, Copilot, medical diagnostics

Introduction

The integration of artificial intelligence (AI) technology into the daily routines and professional workflows of oncologists is steadily increasing. To effectively incorporate AI into clinical practice, it

is crucial to understand the processes involved in the development, validation, and continuous improvement of these technologies.

Interpreting histopathology slides is a complex process that requires specialized skills and in-depth experience. In this process, the pathologist must be able to identify and analyze the different types of tissues and cells present in the slide. Each type of tissue and cell has unique morphological characteristics, which require an in-depth understanding of anatomy and pathology[1, 2]. Along the same lines, interpreting scientific figures published in scientific journals or textbooks can also be challenging for students and laypeople. These figures often contain highly technical and complex information, such as cellular biological mechanisms, mechanisms of tumor development, or cell death pathways, which require a deep understanding of the relevant disciplines[3]. This difficulty is compounded by using specialized terminology and unfamiliar symbols, which can confuse and hinder proper understanding. Without an academic background or experience in these fields, students and laypeople may struggle to understand the context and meaning of these images.

To overcome these challenges, AI-LLM can offer significant assistance. With the ability to understand and analyze language and images, AI-LLM can provide simpler and more accessible explanations of complex scientific figures. AI can identify key elements in an image, translate them into easy-to-understand language, and provide an explanation[4, 5]. Moreover, AI-LLM can offer interactive assistance, where users can ask specific questions and receive relevant answers and feedback, enhancing understanding and making scientific information more accessible to students and the public.

In this study, we will investigate ChatGPT4, Gemini Advanced, and Copilot's ability to interpret histopathology slides and illustrative scientific figures. It discusses the potential of AI LLM's impact in enhancing the accessibility of medical and scientific material to a wider audience. As the amount of visual data used in medical diagnostics and scientific research grows, the capacity to efficiently understand and transmit this information becomes increasingly important. We anticipate that our research will demonstrate how AI-LLM technology can be a valuable tool in assisting specialists in the medical and scientific fields as well as individuals without specialized backgrounds to interpret complex material. This study may help bridge the knowledge gap and open up new opportunities for using AI in different research fields.

Methods

Ethics

Our study was exempt from ethics committee approval because no patients received treatment or intervention, and no information could be linked to a specific patient.

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/). The histopathology images used in part.1 of this study have been approved by the Department of Anatomical Pathology and the Department of Anatomy, Histology and Pharmacology, Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia. The 5 scientific figures used in part.2 are taken from published articles [6-11], which are open to use according to the publishing license.

Study Design

This study was divided into two parts: 1. Histopathology figures interpretation and 2. Scientific figures interpretation. We are testing three AI-LLMs (hereinafter referred to as chatbots): ChatGPT-4 (hereinafter referred to as CG-4), Gemini Advance (hereinafter referred to as GemAdv), and Copilot with various figures (Fig. 1).

In the first part, we will provide 6 figures of histopathology slides consisting of three normal tissue figures (peripheral nerves, kidney, and bone) and three ovarian neoplastic tissue figures (Fig. 2A). Each image will be given a code prompt, and then tested on each chatbot (Supp. 1).

In the second part, we will provide 6 scientific figures (Fig. 2B) consisting of 1)“The role of B7H3 regulating glucose metabolism”[6], 2)“HSF1 plays an important role in tumor cell survival, poor prognosis, and metastasis through several mechanisms”[7], 3)“Carboplatin, HSF1 and Autophagy”[8], 4)” The heat shock response and the regulator of HSF1”[9], 5)” Mechanisms of MST4 induced tumor progression and treatment resistance”[10], and 6)” Three hallmarks of ferroptosis”[11]. Each image will be assigned a code prompt and examined with each chatbot (Supp. 2) (Fig. 1).

The interpretations were promptly recorded into a database, coded, and blindly reviewed by a rater team (4 experts in histopathology in part 1 and 4 lab members who usually work in the field of biomedical science in part 2). To eliminate bias, the AI chatbot responses were coded and randomized before being scored by the raters. The raters evaluated the responses without knowing which came from the chatbot. To analyze the output of the chatbots, we used 5 parameters including "relevance", "clarity", "depth", "focus", and "coherence" [12-15] with a 5-point Likert scale (Tab. 1) [16-18].

Statistical analysis

We investigated the performance of three chatbots in interpreting images. For the evaluation of five parameters, the scores for the five parameters were categorized as follows: 1-2, 3, and 4-5 were classified as "poor," "fair," and "good," respectively [18]. To enhance the interpretability of responses[19-22], the 5-point Likert scale ratings were linearly converted to a 0–100 scale, with higher scores representing superior performance. To assess the consistency among different raters, we reported Pearson and Spearman correlation coefficients for all ratings and employed a one-way ANOVA test with Scheffe's post hoc analysis to examine differences in total scores across raters. The

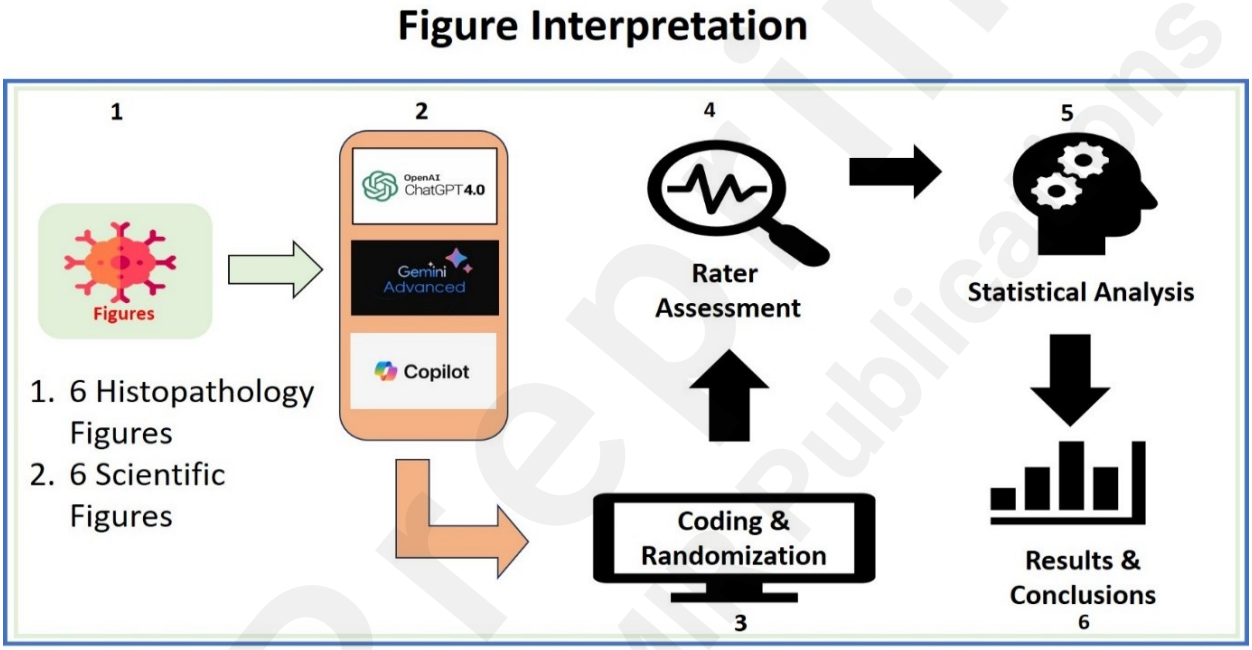

Figure 1. Workflow of image interpretation by chatbots. 6 histopathology and 6 scientific images were tested on each chatbot (CG-4, GemAdv, and Copilot). The interpretation of each chatbot will be coded and randomized to avoid bias. Raters will evaluate all answers, and then statistical analysis will be done.

presence of high inter-rater consistency enhances the generalizability of the evaluation results. To compare the interpretative abilities of three chatbots regarding histopathology images or scientific illustrations, a one-way ANOVA test with Scheffe's post hoc analysis was employed. Furthermore, to reduce the confounding impact of evaluator subjectivity and the complexity of images, multiple linear regression models were utilized. All statistical analyses were performed using the SAS software (Version 9.4, SAS Institute, Cary, NC, USA), with a significance level set at 0.05.

Results

CG-4 surpasses GemAdv and Copilot in providing histopathology image interpretation

Recently, advanced software tools and platforms in digital pathology have emerged to assist pathologists in analyzing histopathological images and making diagnoses [23]. However, not all pathologists, especially those in less developed countries/regions, have access to these smart pathology AI technologies. Therefore, it is crucial to evaluate the value of publicly accessible AI chatbots for interpreting histopathology images. To evaluate the capacities of the chatbots for histopathology image interpretation, 6 histopathology images were tested on each chatbot. The image interpretations by the chatbots were coded and blindly evaluated by four board-certified pathologists (raters). To validate our results, we used three statistical methods to analyze the homogeneity of the raters' rating scores. We found that most raters showed correlation coefficients between 0.82-0.90 by the Pearson test (Fig. 3A), and values between 0.77-0.87 by the Spearman test (Fig. 3B). Furthermore, a one-way ANOVA test with Scheffe's post hoc analysis also showed no significant variation neither (Fig. 3C). All these results indicate that the raters' scores are highly homogenous, indicating that the scoring process is reliable. To assess the quality of the image interpretations by the chatbots, we analyzed the scores of the chatbots in the five individual parameters. Overall, among the three chatbots, CG-4 scored higher than GemAdv and Copilot (Fig. 3D) by a significant (P<.001), suggesting that CG-4 provides better interpretations in providing histopathology explanation (Fig. 3E).

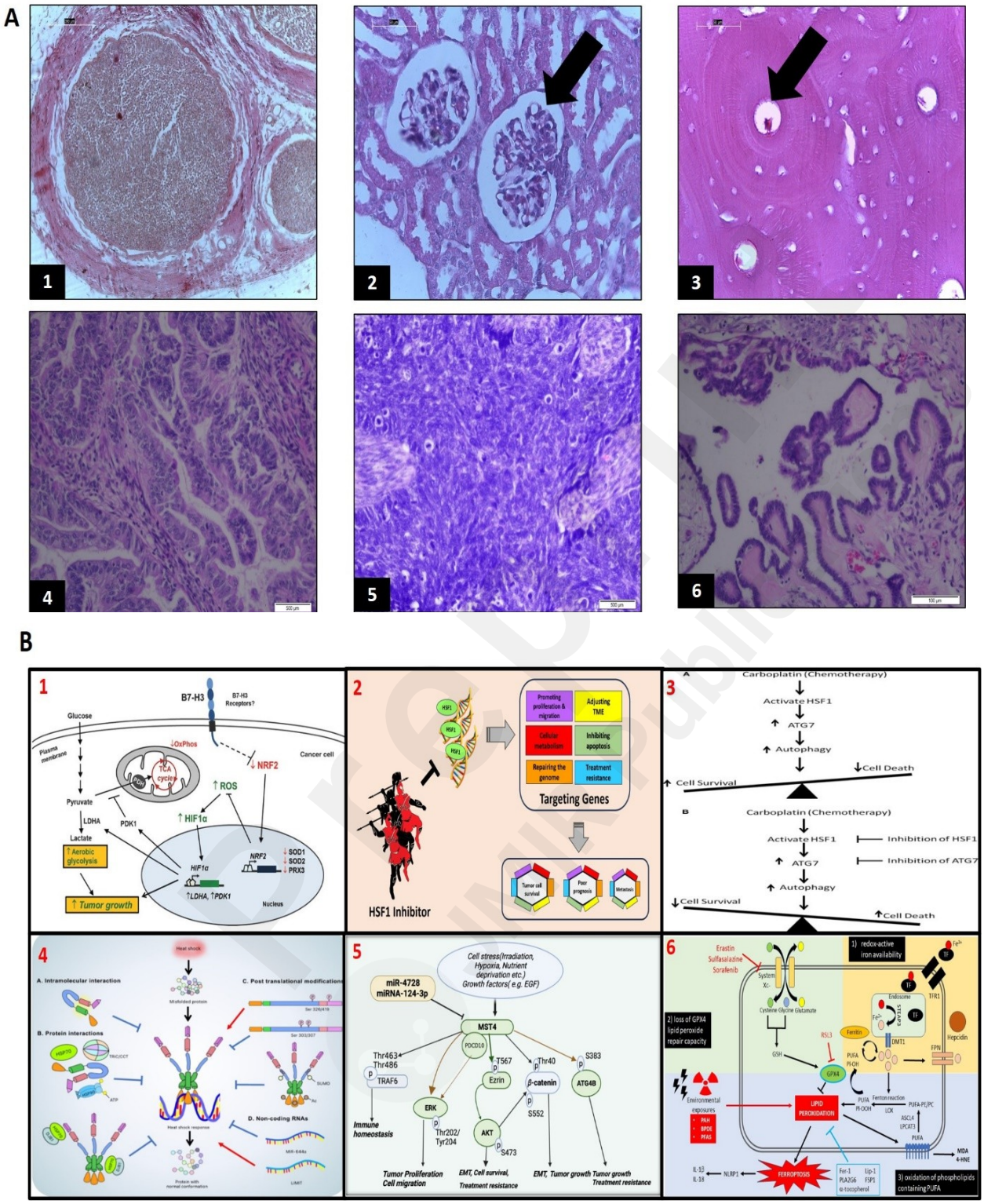

Figure 2. A. Histopathology Slides, consisting of 1) peripheral nerves; 2) kidney; 3) bone; 4) ovarian cancer, endometrioid grade-I; 5) ovarian cancer, endometrioid grade-III; and 6) serous borderline ovarian tumor with implantation.

B. Illustrative scientific figures, consisting of 1) The role of B7H3 regulating glucose metabolism; 2) HSF1 plays an important role in tumor cell survival, poor prognosis, and metastasis through several mechanisms; 3) Carboplatin, HSF1 and Autophagy; 4) The heat shock response and the regulator of HSF1; 5)Mechanisms of MST4 induced tumor progression and treatment resistance; and 6)Three hallmarks of ferroptosis.

Table 1. Assessment parameters and scoring

Assessment Parameters	Definition
Relevance	The response is closely related or appropriate to the issue at hand
Clarity	Clear, easy to understand, free from ambiguity, and transparent
Depth	The answer provides detailed and specific information, not just a general or surface 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 Scale	**Definition**
1 = Very Poor	The answer does not fulfil the basic criteria, is highly irrelevant, or has no effort is evident
2 = Poor	The answer fulfils very few of the expected criteria, with many basic errors
3 = Average	The answer fulfils the basic criteria but does not show more effort or understanding than expected
4 = Good	The answer fulfils all the basic criteria well and shows some aspects that are more than expected
5 = Outstanding	A perfect answer of flawless quality, showing exceptional understanding and complete mastery of the material

CG-4 can provide superior scientific figure interpretation than GemAdv or Copilot

Understanding the figures in scientific papers and textbooks is essential for grasping the key points in the articles, so it is crucial to comprehend them. Next, we tested 6 scientific figures from scientific journals on the 3 chatbots (CG-4, GemAdv, and Copilot). The image interpretations by the chatbots were coded and blindly evaluated by four researchers with intensive training in the fields that are related to scientific figures. To validate our results, we analyzed the rater's scoring homogeneity as aforementioned. We found that most raters showed correlation coefficients between 0.33-0.65 using the Pearson test (Fig. 4A), and values of 0.24-0.72 using the Spearman test (Fig. 4B). Moreover, we used a one-way ANOVA test with Scheffe's post hoc analysis to assess the raters' scoring homogeneity (Fig. 4C). The analyses show that there are no significant statistical variations among these raters, which supports the reliability of our results.

To determine the chatbot that provided the most effective responses, we analyzed the scores of three different chatbots across five parameters. Our findings revealed that CG-4 achieved a significantly higher score than both GemAdv and Copilot (Fig. 4D) by a substantial margin (P<.001). This indicates that CG-4 excels in providing in-depth and coherent explanations of scientific figures. Notably, CG-4 demonstrated superior performance in every aspect, with varying levels of significance. Specifically, it outperformed in Focus and Depth (P<.001), Clarity and Coherence

(P<.01), and Relevance (P<.05) (Fig. 4E), demonstrating its overall strength across various important factors.

Discussion

In this study, significant differences were found in the ability of the three AI-LLMs (ChatGPT-4, Gemini Advanced, and Copilot) to interpret histopathology and scientific figures. CG-4 performed significantly better than GemAdv and Copilot in the first part of the study, which focused on histopathology image interpretation. This was evident from higher average scores on five evaluation parameters, namely "relevance", "clarity", "depth", "focus", and "coherence". The notable variation is probably a result of the CG-4 model being more advanced and well-versed in medical image analysis than the other two AI-LLMs. This conclusion is further supported by the raters' high consensus, showing agreement among the experts in their assessments.

In the second part of the study, CG-4 showed higher proficiency in analyzing scientific figures sourced from academic journals. Despite the lower homogeneity among raters for scoring the interpretation of the scientific figures compared to histopathology images, CG-4 outperformed other chatbots. CG-4's performance showed superiority in all evaluation parameters, with varying levels of significance. This shows that CG-4's ability to understand and convey complex information in scientific images is better than its competitors.

Possible factors that might cause the difference in results between these three AI-LLMs are A) Algorithm and Model Architecture[24]: CG-4 may have used more sophisticated algorithms and model architectures trained on more diverse and specific data, thus providing more accurate and in-depth interpretations. B) Training Data[25, 26]: CG-4 may be trained with larger and more diverse datasets, including histopathology data and scientific images, giving the model a better understanding of different types of images and their context. C) Focus on Medical and Scientific[27]: CG-4 may have components or modules specifically developed for medical and scientific applications, allowing the model to provide more focused and relevant interpretations in these domains. D) User Experience[28]: The wider use of CG-4 and more user feedback may also contribute to the improved performance of this model compared to GemAdv and Copilot.

The participation of histopathology experts and researchers as the rating team in this study had a positive impact and added objectivity. Both groups of raters showed moderate to high homogeneity and consistency, indicating they had similar assessments and perceptions of the interpretations provided by the three AI-LLMs. Additionally, we utilized the five standardized scoring parameters to evaluate the chatbots' responses. Our previous research has demonstrated that this approach

comprehensively evaluates correct and accurate responses [18].

Implications of AI-LLM for patient care

The incorporation of AI technology into the professional practices of oncologists is progressively advancing. A thorough understanding of the methodologies underlying the design, validation, and ongoing optimization of these technologies is essential for their effective integration into clinical practice. Our findings indicate that CG-4 has strong abilities in interpreting histopathology images, which could lead to important benefits for patient care. First, CG-4 can help pathologists make more accurate diagnoses by better identifying and analyzing tissue and cell structures on histopathology slides. This can lower the chances of misdiagnosis and improve overall diagnostic accuracy. Additionally, CG-4's ability to provide quick and accurate interpretations shortens the time needed for diagnosis, which is especially valuable when fast decisions are needed.

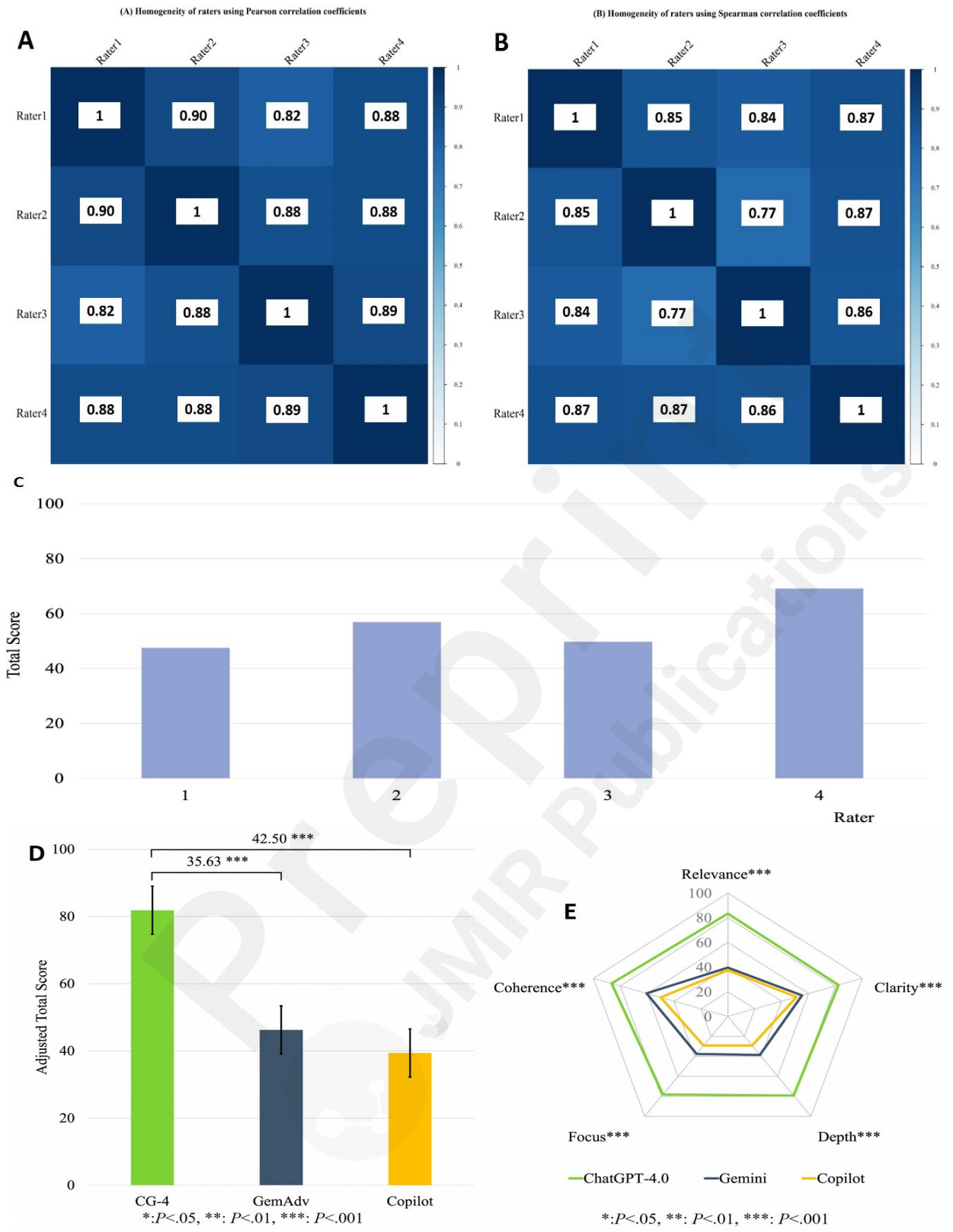

Figure 3. CG-4 surpasses GemAdv and Copilot in providing histopathology image interpretation. Pearson test (**A**) and Spearman test (**B**) showed high homogeneity among raters, respectively 0.82-0.90 & 0.77-0.87. One-way ANOVA test with Scheffe's post hoc analysis showed no significant differences among the raters in giving assessments (**C**). CG-4 scores significantly better than the other two AI-LLMs in providing histopathology image interpretation (**D**). The 5 parameters used for interpretation quality assessment show CG-4's superiority significantly (**E**).

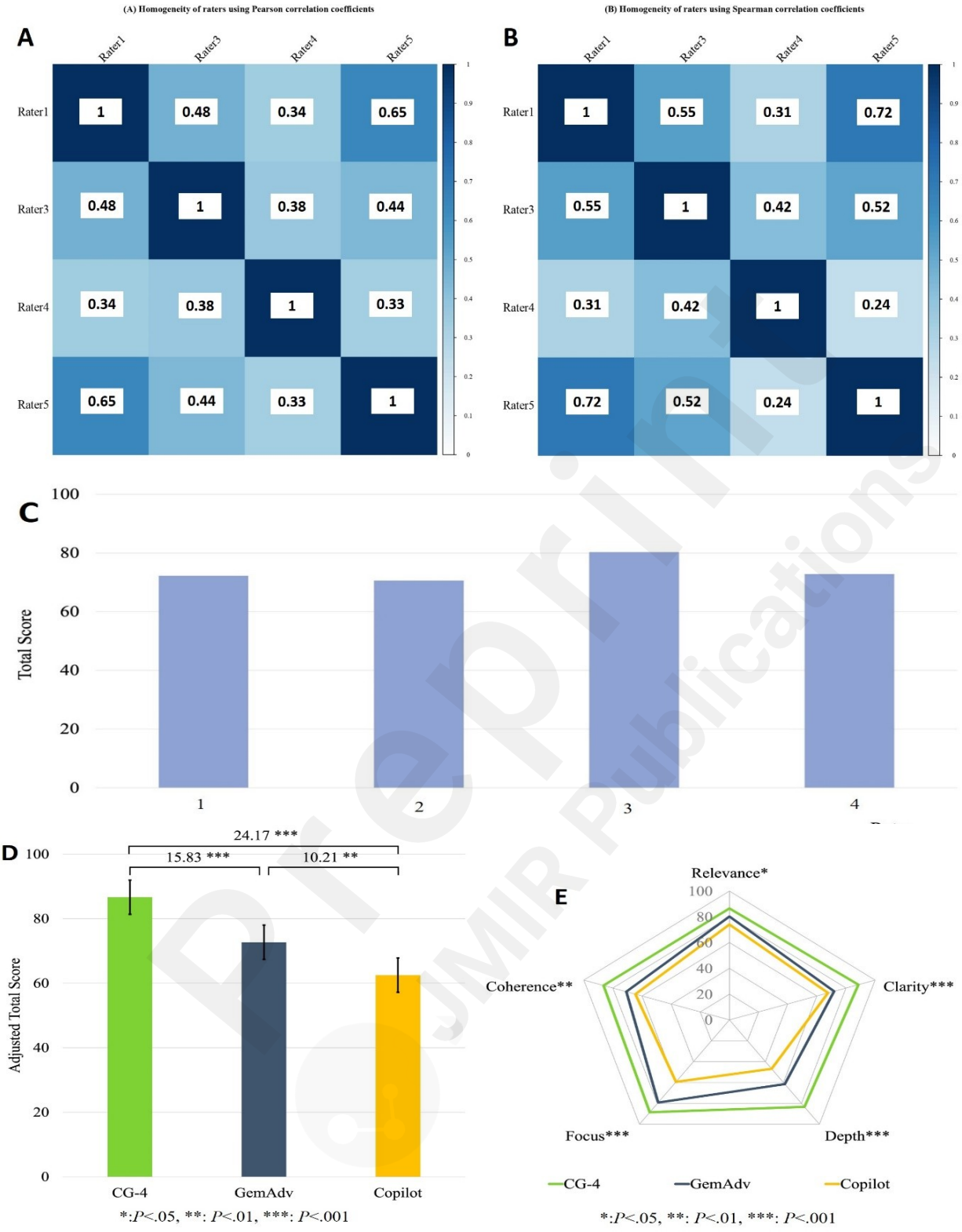

Figure 5. CG-4 can provide superior scientific figure interpretation. Pearson test (**A**) and Spearman test (**B**) showed moderate-high homogeneity among raters, respectively 0.33-0.65 & 0.24-0.72. One-way ANOVA test with Scheffe's post hoc analysis showed no significant differences among the raters in giving assessments (**C**). CG-4 scores significantly better than the other two AI-LLMs in providing scientific image interpretation (**D**). CG-4 convincingly outperformed in 5 assessment parameters with varying significance. Depth and Focus (*P*<.001); Clarity and Coherence (*P*<.01); and relevance (*P*<.05), respectively (**E**).

CG-4 can significantly support pathologists and medical teams by enhancing clinical decision-making and reducing workload. By providing relevant and in-depth information, CG-4 enables pathologists to make better clinical decisions based on accurate data. Additionally, the use of AI-LLMs can reduce the workload of pathologists, allowing them to concentrate on more complex cases that require greater human intervention. Furthermore, CG-4 can improve diagnostic accessibility in remote areas and regions with a shortage of specialist pathologists. In such areas, CG-4 can be an invaluable tool, assisting local medical personnel in making initial diagnoses and offering treatment recommendations.

Impact of AI-LLM on Student Learning of Complex Scientific Figures

The study indicates that CG-4 excels in interpreting scientific figures compared to the other two AI-LLMs, which has several important implications for student learning. CG-4 can enhance students' understanding of complex scientific images by providing clearer, more relevant, and in-depth explanations, thereby facilitating the learning process and improving their comprehension of the material. Additionally, CG-4 promotes learning interactivity, allowing students to engage directly with the AI, ask specific questions, and receive relevant answers, which supports their active and independent learning efforts.

CG-4 also contributes to the development of analytical skills among students by offering opportunities for data analysis exercises and immediate feedback. Students can independently analyze scientific figure data and then verify their results with CG-4, thereby honing their critical analytical skills in the scientific field. The AI provides immediate feedback, helping students identify strengths and weaknesses in their interpretations and guiding them toward skill improvement. Additionally, CG-4 broadens access to learning resources by providing explanations and interpretations of scientific images from various sources that might have been previously inaccessible. This enhances opportunities for broader and deeper learning. Lecturers and researchers can also utilize CG-4 as a teaching and research tool, offering additional explanations that students or research participants may require.

Conclusion

The research delves into examining AI-LLMs for interpreting histopathology slides and scientific figures. The result highlights the powerful capabilities of CG-4 in interpreting both histopathology and scientific figures compared to GemAdv and Copilot. CG-4's advanced performance is evident through higher scores across various evaluation parameters, demonstrating its ability to provide

more accurate, clear, in-depth, focused, and coherent interpretations. The high homogeneity among expert raters further validates these findings.

Several factors likely contribute to CG-4's enhanced performance, including its sophisticated algorithms, diverse and extensive training data, specialized modules for medical and scientific applications, and extensive user feedback. These attributes enable CG-4 to support pathologists and medical teams effectively by improving diagnostic accuracy, reducing the risk of misdiagnosis, enhancing clinical decision-making, and reducing workload. Moreover, CG-4 can provide crucial diagnostic support in remote areas lacking specialist pathologists.

In the educational context, CG-4 offers significant benefits for student learning by enhancing understanding of complex scientific images, promoting interactive learning, and developing analytical skills. It broadens access to learning resources and serves as a valuable tool for lecturers and researchers. Overall, the study underscores the potential of AI-based chatbots to improve patient care through accurate and efficient diagnosis and to enrich student learning in scientific disciplines, making complex information more accessible and understandable.

Acknowledgments

Authorship Contributors

KEG and MT contributed to the conception of the study, methodology, and study design.
KEG, LNL, YCH, and ZYY contributed to data curation and validation.
LNL, YCH, and ZYY contributed to the statistical analysis of the data.
KEG, MT, and LNL contributed to visualization which includes figures, charts, and tables of the data.
KEG and ASR contributed to project administration and resources.
KEG, GA, PAW, R, THY, CH, IHI, CHL, TYH, and MT contributed to question testing and validation.
KEG, DA, EGD, JYY, LNL, and MT contributed to the writing and revising of the manuscript.
KEG and MT were responsible for the decision to submit the manuscript.
Supervision of this research, which includes responsibility for the research activity planning and execution, was overseen by MT.
All authors read and approved the final version of the manuscript.

Data sharing statement

We have ensured that all important data required has been included in the Supplementary file. The exception is the raw values provided by individual doctors, which can be provided upon request.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT-4.0 and Grammarly to edit and proofread the manuscript to improve readability. After using this tool/service, the authors reviewed, verified, and edited the content as needed. The authors take full responsibility for the content of the publication.

Declaration of interests

All authors declare no conflict of interests

Funding

This research was partly funded by the China Medical University Ying-Tsai Scholar Fund CMU109-YT-04, CMU internal fund CMU112-IP-01, and the National Science and Technology Council NSTC 113-2314-B-039-067- (to MT). NSTC 112-2314-B-039 -027 – and 113-2314-B-039 -066 – (to JYY). KEG is a recipient of an Elite Program Scholarship from the Taiwan Ministry of Education.

REFERENCES

1. Cooper M, Ji Z, and Krishnan RG, *Machine learning in computational histopathology: Challenges and opportunities*. Genes Chromosomes Cancer, 2023. **62**(9): p. 540-56.
2. Tommasino C, Merolla F, Russo C, Staibano S, and Rinaldi AM, *Histopathological Image Deep Feature Representation for CBIR in Smart PACS*. J Digit Imaging, 2023. **36**(5): p. 2194-209.
3. Stork DG, *Automatic Computation of Meaning in Authored Images Such as Artworks: A Grand Challenge for AI*. J. Comput. Cult. Herit., 2022. **15**(4): p. Article 65.
4. Minssen T, Vayena E, and Cohen IG, *The Challenges for Regulating Medical Use of ChatGPT and Other Large Language Models*. JAMA, 2023. **330**(4): p. 315-6.
5. Mesko B, *The ChatGPT (Generative Artificial Intelligence) Revolution Has Made Artificial Intelligence Approachable for Medical Professionals*. J Med Internet Res, 2023. **25**: p. e48392.
6. Lim S, Liu H, Madeira da Silva L, Arora R, Liu Z, Phillips JB, Schmitt DC, Vu T, McClellan S, Lin Y, et al., *Immunoregulatory Protein B7-H3 Reprograms Glucose Metabolism in Cancer Cells by ROS-Mediated Stabilization of HIF1alpha*. Cancer Res, 2016. **76**(8): p. 2231-42.
7. Gumilar KE, Chin Y, Ibrahim IH, Tjokroprawiro BA, Yang JY, Zhou M, Gassman NR, and Tan M, *Heat Shock Factor 1 Inhibition: A Novel Anti-Cancer Strategy with Promise for Precision Oncology*. Cancers (Basel), 2023. **15**(21).
8. Desai S, Liu Z, Yao J, Patel N, Chen J, Wu Y, Ahn EE, Fodstad O, and Tan M, *Heat shock factor 1 (HSF1) controls chemoresistance and autophagy through transcriptional regulation of autophagy-related protein 7 (ATG7)*. J Biol Chem, 2013. **288**(13): p. 9165-76.
9. Chin Y, Gumilar KE, Li XG, Tjokroprawiro BA, Lu CH, Lu J, Zhou M, Sobol RW, and Tan M, *Targeting HSF1 for cancer treatment: mechanisms and inhibitor development*. Theranostics, 2023. **13**(7): p. 2281-300.
10. Arora R, Kim JH, Getu AA, Angajala A, Chen YL, Wang B, Kahn AG, Chen H, Reshi L, Lu J, et al., *MST4: A Potential Oncogene and Therapeutic Target in Breast Cancer*. Cells, 2022. **11**(24).
11. Gumilar KE, Priangga B, Lu CH, Dachlan EG, and Tan M, *Iron metabolism and ferroptosis: A pathway for understanding preeclampsia*. Biomed Pharmacother, 2023. **167**: p. 115565.
12. Gordon EB, Towbin AJ, Wingrove P, Shafique U, Haas B, Kitts AB, Feldman J, and Furlan A, *Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions*. J Am Coll Radiol, 2023.
13. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, and Bedayat A, *How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard*. Radiology, 2023. **307**(5): p. e230922.
14. Wu T, He S, Liu J, Sun S, Liu K, Han Q-L, and Tang Y, *A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development*. IEEE/CAA Journal of Automatica Sinica, 2023.

10(5): p. 1122-36.
15. Bhardwaz S and 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. 673-9.
16. Sikander B, Baker JJ, Deveci CD, Lund L, and 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): p. e49019.
17. Veras M, Dyer JO, Rooney M, Barros Silva PG, Rutherford D, and 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**: p. e51873.
18. Gumilar KE, Indraprasta BR, Hsu Y-C, Yu Z-Y, Chen H, Irawan B, Tambunan Z, Wibowo BM, Nugroho H, Tjokroprawiro BA, et al., *Disparities in medical recommendations from AI-based chatbots across different countries/regions*. Scientific Reports, 2024. **14**(1).
19. Daniel C. Ma M, Singh A, Beatrice Bloom M, Nilda Adair R, William Chen M, Husneara Rahman P, Louis Potters M, and Bhupesh Parashar M, DrPH, *Patient Experience Performance at a Primary Cancer Center Versus Affiliated Community Facilities*. Advances in Radiation Oncology, 2023. **8**(5).
20. Kapoor N, Haj-Mirzaian A, Yan HZ, Wickner P, Giess CS, Eappen S, and Khorasani R, *Patient Experience Scores for Radiologists: Comparison With Nonradiologist Physicians and Changes After Public Posting in an Institutional Online Provider Directory*. American Journal of Roentgenology, 2022. **219**(2): p. 338-45.
21. Vaidya TS, Mori S, Dusza SW, Rossi AM, Nehal KS, and Lee EH, *Appearance-related psychosocial distress following facial skin cancer surgery using the FACE-Q Skin Cancer.* Arch Dermatol Res, 2019. **311**(9): p. 691-6.
22. Kamo N, Dandapani SV, Miksad RA, Houlihan MJ, Kaplan I, Regan M, Greenfield TK, and Sanda MG, *Evaluation of the SCA instrument for measuring patient satisfaction with cancer care administered via paper or via the Internet*. Ann Oncol, 2011. **22**(3): p. 723-9.
23. Escobar Díaz Guerrero R, Carvalho L, Bocklitz T, Popp J, and Oliveira JL, *Software tools and platforms in Digital Pathology: a review for clinicians and computer scientists*. J Pathol Inform, 2022. **13**: p. 100103.
24. Taye MM, *Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions*. Computers, 2023. **12**(5).
25. Bazargani R, Fazli L, Gleave M, Goldenberg L, Bashashati A, and Salcudean S, *Multi-scale relational graph convolutional network for multiple instance learning in histopathology images*. Med Image Anal, 2024. **96**: p. 103197.
26. Jenke AC, Bodenstedt S, Kolbinger FR, Distler M, Weitz J, and Speidel S, *One model to use them all: training a segmentation model with complementary datasets.* Int J Comput Assist Radiol Surg, 2024. **19**(6): p. 1233-41.
27. Bitri R and Ali M. *A Comparative Review of GPT-4's Applications in Medicine and High Decision Making*. in *2023 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA)*. 2023.
28. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, Jurafsky D, Szolovits P, Bates DW, Abdulnour RE, et al., *Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study.* Lancet Digit Health, 2024. **6**(1): p. e12-e22.