Comparing the Performance of ChatGPT, GPT-4, and Bard in the Korean Medical Licensing Examination: A Gastroenterology Perspective

Seong Ji Choi, Jeongkuk Seo, Sung Wook Hwang

Submitted to: JMIR Medical Education
on: August 04, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Comparing the Performance of ChatGPT, GPT-4, and Bard in the Korean Medical Licensing Examination: A Gastroenterology Perspective

Seong Ji Choi[1, 2*] MD; Jeongkuk Seo[3*] MD; Sung Wook Hwang[4] MD

[1]Department of Internal Medicine Hanyang University College of Medicine Seoul KR
[2]Department of Internal Medicine Korea University College of Medicine Seoul KR
[3]Department of Internal Medicine Chung-Ang University College of Medicine Seoul KR
[4]Department of Gastroenterology and Inflammatory Bowel Disease Center Asan Medical Center University of Ulsan College of Medicine Seoul KR
[*]these authors contributed equally

Corresponding Author:
Sung Wook Hwang MD
Department of Gastroenterology and Inflammatory Bowel Disease Center
Asan Medical Center
University of Ulsan College of Medicine
East Bldg, 4th Fl.
Seoul
KR

Abstract

Background: Artificial intelligence (AI) and deep learning are revolutionizing the field of medicine, including gastroenterology. Notable among these advancements are large language models (LLMs) like ChatGPT and Bard. While their effectiveness in the United States Medical Licensing Examination has been evaluated, their performance in non-English tests, such as the Korean Medical Licensing Examination (KMLE), remains underexplored.

Objective: This study examined the performance of LLMs in the KMLE and assessed the explanation quality for gastroenterology questions.

Methods: We analyzed the performance of three LLMs, ChatGPT (GPT-3.5), GPT-4, and Bard in the official KMLE questions from 2021 and 2022 by focusing on their accuracy and consistency. The questions were categorized based on their subject and the inclusion of images. The quality of explanations provided by the LLMs for gastroenterology questions was assessed using the Global Quality Score (GQS).

Results: In KMLE 2021 and 2022, GPT-4 achieved accuracy rates of 72.1% and 62.5%, respectively, with consistency of approximately 60%. These rates were significantly higher than those of ChatGPT and Bard. The higher accuracy was maintained across most subjects, as well as in questions with images. For gastroenterology questions, GPT-4 achieved accuracy rates of 55.8% and 63.9% in KMLE 2021 and 2022, respectively. The GQS of GPT-4 was significantly higher than those of other models, even in cases of incorrect responses.

Conclusions: Unlike ChatGPT and Bard, GPT-4 achieved passing grades for the KMLE 2021 and 2022, and demonstrated high-quality explanations in gastroenterology. Further research on multimodal models and the use of prompt engineering is warranted.

(JMIR Preprints 04/08/2024:65063)
DOI: https://doi.org/10.2196/preprints.65063

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

Comparing the Performance of ChatGPT, GPT-4, and Bard in the Korean Medical Licensing Examination: A Gastroenterology Perspective

Running title: LLM Performance in KMLE: ChatGPT vs. GPT-4 vs. Bard

Seong Ji Choi[1,2*], Jeongkuk Seo[3,*], Sung Wook Hwang[4]

[1]Department of Internal Medicine, Hanyang University College of Medicine, Seoul, Korea

[2]Department of Internal Medicine, Korea University College of Medicine, Seoul, Korea

[3]Department of Internal Medicine, Chung-Ang University College of Medicine, Seoul, Korea

[4]Department of Gastroenterology and Inflammatory Bowel Disease Center, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

*Seong Ji Choi and Jeongkuk Seo equally contributed as a first author.

Corresponding author: Sung Wook Hwang, MD, PhD

Department of Gastroenterology and Inflammatory Bowel Disease Center

Asan Medical Center, University of Ulsan College of Medicine

88, Olympic-ro 43-gil, Songpa-gu, Seoul 05505, South Korea

Tel: +82-2-3010-3195; Fax: +82-2-476-0824; Email: hsw903@gmail.com

ABSTRACT

Background/Objectives: Artificial intelligence (AI) and deep learning are revolutionizing the field of medicine, including gastroenterology. Notable among these advancements are large language models (LLMs) like ChatGPT and Bard. While their effectiveness in the United States Medical Licensing Examination has been evaluated, their performance in non-English tests, such as the Korean Medical Licensing Examination (KMLE), remains underexplored. This study examined the performance of LLMs in the KMLE and assessed the explanation quality for gastroenterology questions.

Methods: We analyzed the performance of three LLMs, ChatGPT (GPT-3.5), GPT-4, and Bard in the official KMLE questions from 2021 and 2022 by focusing on their accuracy and consistency. The questions were categorized based on their subject and the inclusion of images. The quality of explanations provided by the LLMs for gastroenterology questions was assessed using the Global Quality Score (GQS).

Results: In KMLE 2021 and 2022, GPT-4 achieved accuracy rates of 72.1% and 62.5%, respectively, with consistency of approximately 60%. These rates were significantly higher than those of ChatGPT and Bard. The higher accuracy was maintained across most subjects, as well as in questions with images. For gastroenterology questions, GPT-4 achieved accuracy rates of 55.8% and 63.9% in KMLE 2021 and 2022, respectively. The GQS of GPT-4 was significantly higher than those of other models, even in cases of incorrect responses.

Conclusions: Unlike ChatGPT and Bard, GPT-4 achieved passing grades for the KMLE 2021 and 2022, and demonstrated high-quality explanations in gastroenterology. Further research on multimodal models and the use of prompt engineering is warranted.

Keyword: ChatGPT; GPT-4; Bard; natural language processing; medical examination

Introduction

The application of artificial intelligence (AI) and deep learning (DL) techniques into the medical field has revolutionized various aspects of patient care and medical data analysis.[1] The field of gastroenterology has also benefited from the capabilities of DL, as its ability to efficiently categorize diverse endoscopic images has significantly improved the diagnosis and detection of many pathological conditions.[2] However, the advancements driven by DL in gastroenterology have been largely confined to image analysis, potentially neglecting other important critical facets of AI. Meanwhile, the increasing use of DL in natural language processing (NLP) has garnered significant attention, particularly with the introduction of transformer architectures and the rise of large language models (LLMs).[3,4] The availability of vast amounts of text data, advancements in hardware such as GPUs, and ongoing research into DL architectures and techniques have resulted in the development of high-performance LLMs such as ChatGPT (Chat Generative Pre-trained Transformer) (OpenAI, San Francisco, CA, USA) and Bard (Google LLC, Mountain View, CA, USA).[5] These models contain over 100 billion parameters and predict the probability of word sequences based on the context of the training data. With adequate training, LLMs can generate responses that mimic human interaction in response to user queries. However, the currently available LLMs differ in many ways, including the number of parameters, training sources, and even pricing policies.[6]

LLMs have been used in various fields and have even been tested in qualifying examination questions to assess their effectiveness. LLMs showed human-level performance on most of these examinations and even achieved a top 10% score on the Uniform Bar Examination.[7] In the field of medicine, LLMs have performed well on many exams, including the US Medical Licensing Examination and various subspecialty examinations such as ophthalmology and cardiology. However, they did not perform well on the American College of Gastroenterology self-assessment test and Taiwan's 2022 Family Medicine Board Exam.[8-12] Similarly, in a study assessing the performance of LLM in pediatric diagnostics, the model's error rate was 83%, indicating significant challenges in accurate medical interpretation, particularly in specialized fields due to their complexity.[13] These findings have sparked interest in evaluating the models' performance in specialized medical fields and their potential utility in new domains, such as non-English medical exams. Furthermore, it is imperative to assess not only the models' performance in these exams, but also the quality of their performance in terms of clinical applicability.

Therefore, we aimed to evaluate the performance of three publicly available language models (LLMs)—ChatGPT, GPT-4, and Bard—in answering official Korean Medical Licensing Examination (KMLE) questions without translation. We also analyzed the quality of responses in the gastroenterology field to assess the potential and limitations of these models in term of medical diagnostics and reasoning.

Methods

Medical Licensing Datasets

Official KMLE questions from 2021 and 2022 were downloaded from the Korea Health Personnel Licensing Examination Institute website with authorized permission.[14] The 2021 exam consisted of 360 questions over five sessions, while the 2022 exam had 320 questions over four sessions, reflecting a change in format for the computer-based test. All questions are in Korean, but units are in English, and some specific nouns are written in both Korean and English. The exam consists of 4 parts: 1) major subjects including Internal Medicine, General Surgery, Obstetrics and Gynecology, Pediatrics, and Psychiatry; 2) minor subjects including Neurology, Neurosurgery, Otorhinolaryngology, Ophthalmology, Dermatology, Emergency Medicine, Orthopedics, and Urology; 3) Preventive Medicine; and 4) Medical Law. The evaluation is graded on an absolute basis, and an overall average score of 60% with no individual session scoring below 40% is needed to pass the exam. For the analysis, questions were labeled based on the inclusion of images and by subject. Gastroenterology questions were further categorized into liver, gastrointestinal tract, or pancreaticobiliary.

Study design

The KMLE questions were posed to three LLMs—ChatGPT (GPT-3.5), GPT-4, and Bard—from August 5th to September 21st, 2023. While ChatGPT and Bard are provided for free, GPT-4 is available through a premium subscription. GPT-4 and Bard have a limit on the number of questions that can be asked within a given time period; therefore, the question and answer process occurred over several weeks. For ChatGPT, we used the August 3 version, and for Bard, we used the August 3 update.

All labeling, categorization, question entry, scoring of response quality, and statistical analysis were performed by three professors of gastroenterology with over 5 years of clinical experience. To minimize the risk of memory retention bias, a new chat window was utilized for each question. The first prompt was, "This question is from the KMLE. Please solve it." in Korean. Following this prompt, we copied and pasted each question into the chat box. We did not conduct any prompt engineering beyond this statement unless the model refused to answer. For the questions that the model refused to answer, we modified the questions so that the model could respond appropriately. For the questions with images, we included all the details except the images. Three gastroenterologists inputted each question into the model and aggregated the answers to arrive at a final response. The final answer was determined by a majority vote, and if all three answers were different, the response was considered incorrect. Consistency was defined as when all three answers were identical.

Analysis of response quality

To evaluate the quality of the responses, we gathered the complete responses of all LLMs to the gastroenterology questions. The order of the three responses to each question was then randomized, the format was standardized, and the results were evaluated by three gastroenterologists in a blinded manner. The quality assessment of the LLMs' response was conducted using the Global Quality Score (GQS), a five-point Likert scale developed by Bernard et al., in which 1 denotes the lowest score and 5 denotes the highest score.[15] This scoring system evaluates the coherence and quality of the description, ensuring the inclusion of essential information, its usefulness to patients, and comprehensive coverage of crucial topics. The final score for each question was calculated by averaging the scores of the three gastroenterologists.

Statistical analysis

Accuracy and consistency are presented as fractions and percentages, calculated by dividing the number of correct results by the total number of questions. GQSs are presented in terms of mean value. Chi-squared analysis and Fisher exact test were performed to compare the performance of the LLMs. The Kruskal-Wallis test was used to compare the GQSs of LLMs on gastroenterology questions, followed by a post-hoc Mann-Whitney U test with the Bonferroni correction method. A p-value less than 0.05 was considered statistically significant. All analyses were performed using IBM SPSS Statistics Version 27.0 (IBM Corp., Armonk, NY, USA).

Ethics

Ethical approval was not required as the study did not involve human or animal subjects.

RESULTS

Overall Performance

The overall performance of three LLMs on the KMLE 2021 and 2022 exams is shown in Table 1. GPT-4 achieved the highest accuracy rates of 72.1% (259/359) and 62.5% (198/317) on all questions from the KMLE 2021 and 2022 exams, respectively, outperforming ChatGPT, which scored 34.8% and 30.3%, and Bard, which scored 54.3% and 50.5%, respectively (all $p < 0.001$). When analyzing based on whether the questions included images, GPT-4 still exhibited the best performance, especially for the questions without images (all $p < 0.001$). In terms of consistency, GPT-4 demonstrated the best results, achieving rates of 64.9% for 2021 and 60.9% for 2022, outperforming ChatGPT and Bard.

Next, the performance of three LLMs was analyzed based on the subjects. In the KMLE questions from 2021, GPT-4 demonstrated the highest accuracy in most subjects, including internal medicine, law, obstetrics and gynecology, pediatrics, preventive medicine, and psychiatry. Meanwhile, GPT-4 and Bard showed similar results in surgery. In the KMLE questions from 2022, GPT-4 showed the highest accuracy in most subjects, except for law, in which Bard demonstrated the best performance with 70.8% accuracy, surpassing ChatGPT (37.5%) and GPT-4 (29.2%).

Performance on Gastroenterology Questions

To evaluate the performance of the three LLMs in a specialized medical field, their accuracy in responding to gastroenterology questions was evaluated (Table 2). GPT-4 achieved the highest accuracy rates of 55.8% (24/43) and 63.9% (23/36) on gastroenterology questions from the KMLE 2021 and 2022, respectively, which was significantly superior to those of ChatGPT, which scored 27.9% and 27.8%, and Bard, which scored 41.9% and 38.9%, respectively (all $p < 0.05$). When assessing based on image inclusion, GPT-4 still demonstrated superior results, although the difference was not statistically significant, except for the 2022 KMLE questions with images, which was likely due to the limited number of questions available. After categorizing the gastroenterology questions into liver, gastrointestinal tract, or pancreaticobiliary parts, GPT-4 still showed better accuracy compared to ChatGPT and Bard.

Response Quality to Gastroenterology Questions

The three gastroenterologists evaluated the quality of the responses from the three LLMs to gastroenterology questions using GQS (Table 3 and Figure 1). The response quality of GPT-4 to gastroenterology questions was the highest among the three LLMs, achieving a score of 3.81 in both KMLE 2021 and 2022 (all $p < 0.001$). GPT-4 had the highest proportion of GQS scores 4 and 5, indicating a superior quality of response to the gastroenterology questions compared to ChatGPT and Bard (Figure 1). The GQS score of GPT-4 was higher for questions without images compared to those with images (Table 3). Additionally, when assessed according to the three subcategories, GPT-4 showed superior performance in the gastrointestinal tract part in 2021 and 2022, and in the liver and pancreaticobiliary parts in 2021.

To evaluate the appropriateness of the LLMs' responses, regardless of their correctness, we further analyzed the GQS based on the accuracy of the responses (Table 4). GPT-4 showed superior GQS scores for gastroenterology questions from KMLE 2021 and 2022, when it provided correct answers. It also maintained better GQS for questions with incorrect answers from KMLE 2021 ($p < 0.001$), showing the application of appropriate reasoning even in instances where the answers were wrong.

DISCUSSION

This study presents a novel examination of the capabilities of LLMs, specifically ChatGPT, GPT-4, and Google Bard, in responding to questions from the KMLE, an official licensing exam for Korean doctors, with an emphasis on gastroenterology. GPT-4 outperformed other models in answering questions, whether they included images or not, and demonstrated superior overall performance across most subjects. Also, the quality of the explanations of GPT-4 for gastroenterology questions was significantly better than those of ChatGPT or Bard. Only GPT-4 achieved an accuracy rate of over 60% in both 2021 and 2022, which is the passing mark of KMLE.

Therefore, our findings reveal significant differences in performance among the LLMs, highlighting the rapid evolution and potential of LLMs in medical education and clinical practice. The performance of the best current LLM, particularly GPT-4, can be compared to that of a newly graduated doctor, indicating that while its potential in the medical field is promising, it should be used with caution and supervision.

Since the introduction of LLMs such as ChatGPT and Bard, numerous studies have evaluated their medical capabilities in specific fields. However, the majority of these studies do not reflect the latest advancements in LLMs, such as GPT-4. In addition, they are often limited to brief reports or letters, assessing only the accuracy of LLMs in medical examinations, resulting in a lack of comprehensive analysis of LLM functionalities.[8,16] Particularly, there is a significant gap in research on inputting medical questions into the LLM between English and non-English languages, and it is important to ensure that LLMs are able to produce acceptable results when medical questions are asked in those languages. Existing research on the accuracy of LLMs in Korean medical examination questions has primarily focused on specialist exams in surgery and dermatology.[17,18] Our study fills this gap by providing a more comprehensive evaluation of LLM performance in the medical field, specifically focusing on a non-English language. In addition, a significant strength of our study is the use of official examination questions from the KMLE. This approach is a noteworthy advantage because many other studies tend to use questions from different sources, which often lack comparability with actual examination content. By utilizing officially approved questions, our research offers a more precise and pertinent evaluation of the LLMs' abilities in a real-world medical examination setting.

The superior performance of GPT-4 in both overall and gastroenterology-specific questions indicates recent advancements in AI's comprehension of complex medical subjects. GPT-4 also demonstrated superiority over other models in terms of response consistency, which makes it a more reliable model for potential use in training or clinical practice. The enhanced reasoning and analytical capabilities of GPT-4 is likely attributable to its fine-tuning, along with the introduction of a rule-based reward model (RBRM) approach and training on a more extensive dataset with a significantly larger architectural model size compared to ChatGPT and Bard.[7,19] Our findings are in line with earlier studies in the fields of neurosurgery and ophthalmology, which emphasized the superiority of GPT-4 over other LLMs.[20-22]

At the time of this study (Fall 2023), models other than GPT-4 did not support image input. Therefore, questions containing images were inputted into these models without the image components. As expected, we observed that in KMLE 2021, the LLMs exhibited higher accuracy in answering questions that initially did not have images compared to those that initially had image components; interestingly however, the results of GPT-4 in KMLE 2022 showed similar accuracies regardless of the presence of images. Nonetheless, when the image components were inputted to the prompt, GPT-4 ironically showed subpar accuracy in answering the questions. This suggests that while textual analysis is advancing, there is a need for future development in the integration and analysis of visual data, especially in the medical field.

When the performance of LLMs was analyzed across different subjects, our findings revealed significant differences in their performance when regional characteristics were reflected in the exam questions, particularly in the context of medical law. In the 2021 exam, the accuracy rates for medical law questions among various LLMs mirrored the overall trend. However, in the 2022 exam, a discrepancy was observed as while GPT-4 showed an accuracy rate of only 29.2%, Bard achieved a significantly higher rate of 70.8%. This difference may indicate variations in the training datasets for each LLM and the timing of their training, suggesting that regional and temporal factors significantly impact LLM performance.

The application of LLMs in medical education and assessment, especially in non-English speaking environments, is a growing area of interest. The capability of these models to not only produce accurate answers but also to generate high-quality explanations in the user's language is especially important in clinical education, where comprehending the reasoning behind medical decisions is crucial. While advancements in LLMs have significantly reduced instances of hallucinations, it remains imperative that users perform cross-verification for the given information. In such situations, the high-quality explanations provided by the LLMs not only improve our comprehension of the material but also enable users to verify the accuracy of the information. In this context, we sought to investigate the quality of explanations of LLMs using the GQS. Our results highlight the importance of not only the accuracy but also the comprehensibility and relevance of AI-generated content in medical settings. When examining the average GQS of LLMs for gastroenterology questions, it was observed that GPT-4 generally outperformed other models. Following this, Bard tended to demonstrate similar or better explanatory capabilities compared to ChatGPT. Furthermore, when analyzing the GQS based on correct versus incorrect answers, it was predictably found that the GQS for incorrect answers was consistently lower than for correct ones. Notably, in the case of questions from the 2021 KMLE, Bard's GQS, when answered correctly, was as high as 3.93, rivaling that of GPT-4. In contrast, for the 2022 questions, when the answers were incorrect, GPT-4's GQS dropped to 2.87, indicating no significant difference from the GQS of ChatGPT or Bard. The differences in explanatory power among LLMs could potentially result from variations in model size or training data. Therefore, it is plausible that the complexity or specificity of a question may result in better performance by a particular model.

This study has the following limitations. The main limitation is its focus on KMLE gastroenterology questions, which may restrict the applicability of the findings to other medical specialties, languages, and healthcare systems. This narrow focus does not fully capture the potential capabilities of LLMs in addressing a wide range of medical questions and problems across various medical fields. Secondly, as AI models are continually updated and improved, newer versions of these models may exhibit different performance characteristics. Thirdly, while three expert gastroenterologists conducted the scoring and assessment of response quality, there is a possibility of subjective bias being introduced. However, the GQS scores were relatively consistent among the three scoring gastroenterologists, and we employed measures such as shuffling the questions, blinding, and averaging the GQS scores to accurately assess the explanatory power of each LLM.

Although our study found that GPT-4 has demonstrated performance near or above the passing mark, using a model that only achieves 60% or 70% accuracy in educational settings or clinical practice would be considered unacceptable. Future research should continue to validate the latest AI models and explore the integration of AI in medical image analysis, an area that remains underdeveloped compared to textual analysis. Furthermore, additional studies could explore the performance of AI models trained specifically for non-English languages and medical systems, such as Naver's HyperCLOVA. These studies would offer a more comprehensive understanding of the capabilities and limitations of AI in the global medical education landscape.

In conclusion, our study highlights the significant potential and accompanying weaknesses of AI models such as ChatGPT and Google Bard in medical education and assessment, especially in the field of gastroenterology. As these models continue to evolve, they promise to be valuable tools in medical training and continuous learning, enhancing the capabilities of medical professionals and potentially transforming the landscape of medical education and clinical practice.

Statement

While preparing this document, the authors used ChatGPT and GPT-4 for the following purposes: (i) to rephrase original sentences for improved readability, (ii) to search for specific information, and (iii) to summarize and clarify text from relevant sources and references. Following the use of this tool, the authors thoroughly reviewed and revised the content as necessary, and assume complete responsibility for the final published material.

Table 1. Performance of LLMs on KMLE 2021 and 2022

	KMLE (2021)				KMLE (2022)			
	ChatGPT	**GPT-4**	**Bard**	**p-value**	**ChatGPT**	**GPT-4**	**Bard**	**p-value**
Accuracy of all questions, n/N (%)	125/359 (34.8)	259/359 (72.1)	195/359 (54.3)	<0.001	96/317 (30.3)	198/317 (62.5)	160/317 (50.5)	<0.001
Questions without images	64/180 (35.6)	145/180 (80.6)	103/180 (57.2)	<0.001	55/149 (36.9)	94/149 (63.1)	82/149 (55.0)	<0.001
Questions with images	61/179 (34.1)	114/179 (63.7)	92/179 (51.4)	<0.001	41/168 (24.4)	104/168 (61.9)	78/168 (46.4)	<0.001
Consistency, n/N (%)	76/359 (21.2)	233/359 (64.9)	151/359 (42.1)	<0.001	54/317 (17.0)	193/317 (60.9)	100/317 (31.5)	<0.001
Accuracy of subjects, n/N (%)								
Internal medicine	53/158 (33.5)	104/158 (65.8)	72/158 (45.6)	<0.001	36/149 (24.2)	92/149 (61.7)	62/149 (41.6)	<0.001
Law	7/19 (36.8)	13/19 (68.4)	8/19 (42.1)	0.11	9/24 (37.5)	7/24 (29.2)	17/24 (70.8)	0.009
Obstetrics and gynecology	17/44 (38.6)	37/44 (84.1)	20/44 (45.5)	<0.001	11/38 (29.0)	22/38 (57.9)	20/38 (52.6)	0.032
Pediatrics	19/52 (36.5)	36/52 (69.2)	35/52 (67.3)	0.001	15/32 (46.9)	24/32 (75.0)	19/32 (59.4)	0.18
Preventive medicine	10/24 (41.7)	20/24 (83.6)	17/24 (70.8)	0.008	9/18 (50.0)	13/18 (72.2)	11/18 (61.1)	0.48
Psychiatry	8/25 (32.0)	22/25 (88.0)	19/25 (76.0)	<0.001	6/18 (33.3)	15/18 (83.3)	10/18 (55.6)	0.040
Surgery	7/24 (29.2)	16/24 (66.7)	16/24 (66.7)	0.011	6/15 (40.0)	8/15 (53.3)	5/15 (33.3)	0.53
Minor	4/13 (30.8)	11/13 (84.6)	8/13 (61.5)	0.020	4/23 (17.4)	17/23 (73.9)	16/23 (69.6)	<0.001

Values are n/N (%), where n = number of correct answers, N = total number of questions, % = percentage calculated as n/N x 100

Table 2. Performance of LLMs on gastroenterology questions

	KMLE (2021)				**KMLE (2022)**			
	ChatGPT	**GPT-4**	**Bard**	**p-value**	**ChatGPT**	**GPT-4**	**Bard**	**p-value**
Accuracy of GEQ, n/N (%)	12/43 (27.9)	24/43 (55.8)	18/43 (41.9)	0.032	10/36 (27.8)	23/36 (63.9)	14/36 (38.9)	0.007
Without images	4/13 (30.8)	10/13 (76.9)	7/13 (53.9)	0.05	3/8 (37.5)	5/8 (62.5)	1/8 (12.5)	0.17
With images	8/30 (26.7)	14/30 (46.7)	11/30 (36.7)	0.27	7/28 (25.0)	18/28 (64.3)	13/28 (46.4)	0.013
Accuracy of parts, n/N (%)								
Hollow	8/24 (33.3)	14/24 (58.3)	12/24 (50.0)	0.21	9/21 (42.9)	14/21 (66.7)	10/21 (47.6)	0.26
Liver	2/10 (20.0)	7/10 (70.0)	5/10 (50.0)	0.10	0/6 (0.0)	6/6 (100.0)	1/6 (16.7)	0.001
PB	2/9 (22.2)	3/9 (33.3)	1/9 (11.1)	0.84	1/9 (11.1)	3/9 (33.3)	3/9 (33.3)	0.63

Values are n/N (%), where n = number of correct answers, N = total number of questions, % = percentage calculated as n/N x 100

GEQ, gastroenterology question; PB, pancreaticobiliary.

Table 3. Global Quality Score (GQS) of LLMs on gastroenterology questions

	KMLE (2021)					KMLE (2022)				
	ChatGPT[A]	**GPT-4**[B]	**Bard**[C]	**p-value**	**Post-hoc***	**ChatGPT**[A]	**GPT-4**[B]	**Bard**[C]	**p-value**	**Post-hoc***
GEQ	2.34	3.81	3.15	<0.001	A<C<B	2.91	3.81	3.00	<0.001	A=C<B
GEQ without images	2.24	4.02	3.14	0.003	A<C<B	3.13	3.96	2.96	0.14	-
GEQ with images	2.39	3.70	3.15	<0.001	A<C<B	2.85	3.76	3.01	0.002	A=C<B
Parts										
GI tract	2.42	3.67	3.17	0.002	A<C<B	2.93	3.90	3.19	0.006	A=C<B
Liver	2.07	4.07	3.27	0.012	A<C<B	3.38	4.17	3.11	0.12	-
PB	2.44	3.88	2.96	0.038	A=C<B	2.52	3.33	2.48	0.26	-

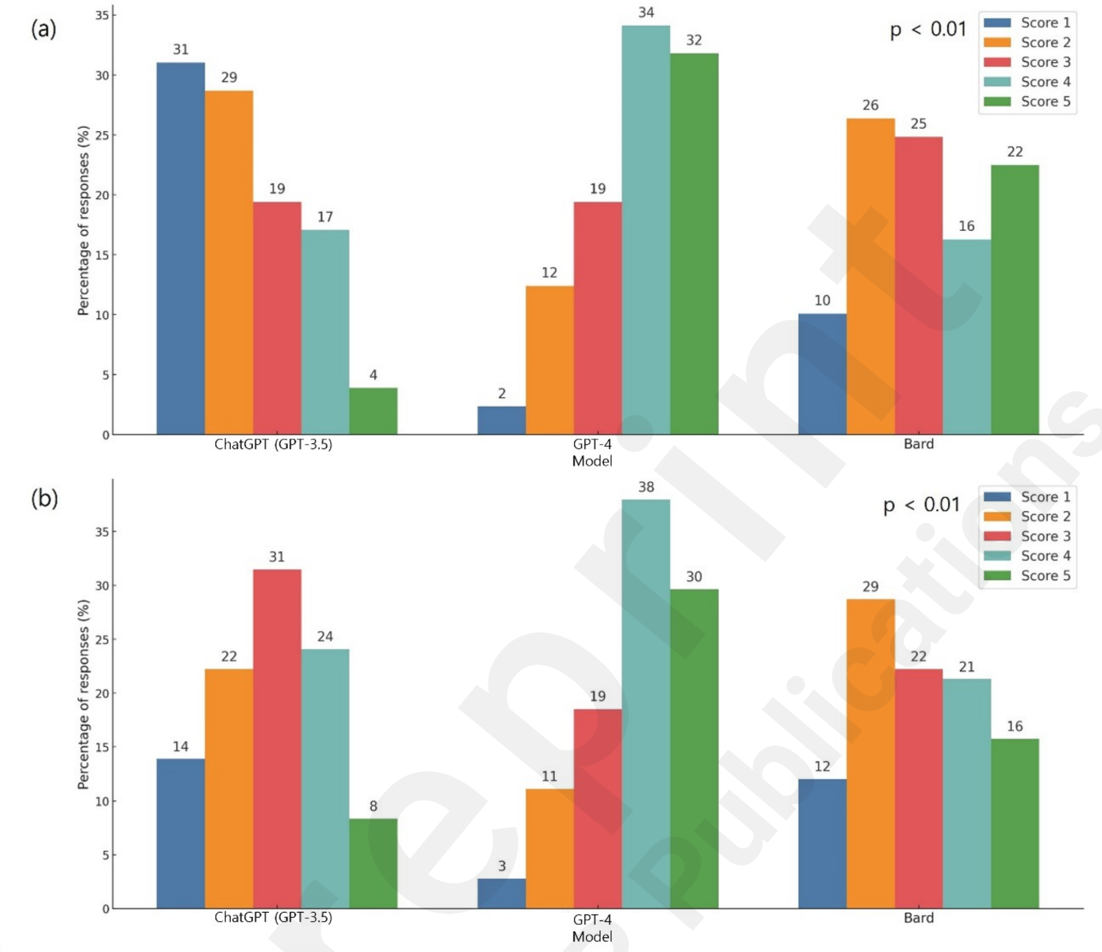
GEQ, gastroenterology question; GI, gastrointestinal; PB, pancreaticobiliary. *The significance level was corrected from 0.05 to 0.017 (5%/3) by the Bonferroni correction method.

Table 4. Effect of correctness on the Global Quality Score (GQS) of LLMs

	KMLE (2021)					**KMLE (2022)**				
	ChatGPT^{A}	**GPT-4^{B}**	**Bard^{C}**	**p-value***	**Post-hoc**	**ChatGPT^{A}**	**GPT-4^{B}**	**Bard^{C}**	**p-value***	**Post-hoc**
Correct answer	3.21	4.18	3.93	0.013	A<B=C	3.40	4.33	3.90	0.004	A<C<B
Wrong answer	1.92	3.41	2.59	<0.001	A<C<B	2.72	2.87	2.42	0.33	-

*The significance level was corrected from 0.05 to 0.017 (5%/3) by the Bonferroni correction method.

Figure 1. Comparison of global quality scores (GQS) of the responses of LLMs to gastroenterology questions. (a) KMLE (2021), (b) KMLE (2022).



REFERENCES

1 Alowais, S. A. *et al.* Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ* **23**, 689 (2023). https://doi.org/10.1186/s12909-023-04698-z

2 Okagawa, Y., Abe, S., Yamada, M., Oda, I. & Saito, Y. Artificial Intelligence in Endoscopy. *Dig Dis Sci* **67**, 1553-1572 (2022). https://doi.org/10.1007/s10620-021-07086-z

3 Vaswani, A. *et al.* Attention Is All You Need. arXiv:1706.03762 (2017). <https://ui.adsabs.harvard.edu/abs/2017arXiv170603762V>.

4 Zhao, W. X. *et al.* A Survey of Large Language Models. arXiv:2303.18223 (2023). <https://ui.adsabs.harvard.edu/abs/2023arXiv230318223Z>.

5 Kevin, S. in *The Official Microsoft Blog.* Vol. 2023 (2020).

6 Ray, P. P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* **3**, 121-154 (2023). https://doi.org/https://doi.org/10.1016/j.iotcps.2023.04.003

7 OpenAI. GPT-4 Technical Report. arXiv:2303.08774 (2023). <https://ui.adsabs.harvard.edu/abs/2023arXiv230308774O>.

8 Suchman, K., Garg, S. & Trindade, A. J. Chat Generative Pretrained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test. *Am J Gastroenterol* (2023). https://doi.org/10.14309/ajg.0000000000002320

9 Weng, T. L., Wang, Y. M., Chang, S., Chen, T. J. & Hwang, S. J. ChatGPT failed Taiwan's Family Medicine Board Exam. *J Chin Med Assoc* **86**, 762-766 (2023). https://doi.org/10.1097/JCMA.0000000000000946

10 Kung, T. H. *et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health* **2**, e0000198 (2023). https://doi.org/10.1371/journal.pdig.0000198

11 Gilson, A. *et al.* How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ* **9**, e45312 (2023). https://doi.org/10.2196/45312

12 Skalidis, I. *et al.* ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? *Eur Heart J Digit Health* **4**, 279-281 (2023). https://doi.org/10.1093/ehjdh/ztad029

13 Barile, J. *et al.* Diagnostic Accuracy of a Large Language Model in Pediatric Case Studies. *JAMA Pediatrics* (2024). https://doi.org/10.1001/jamapediatrics.2023.5750

14 Institute, K. H. P. L. E. <kuksiwon.or.kr> (2023).

15 Bernard, A. *et al.* A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am J Gastroenterol* **102**, 2070-2077 (2007). https://doi.org/10.1111/j.1572-0241.2007.01325.x

16 Fijačko, N., Gosak, L., Štiglic, G., Picard, C. T. & Douma, M. J. Can ChatGPT pass the life support exams without entering the American heart association course? *Resuscitation* (2023). https://doi.org/https://doi.org/10.1016/j.resuscitation.2023.109732

17 Oh, N., Choi, G. S. & Lee, W. Y. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and

training in the era of large language models. *Ann Surg Treat Res* **104**, 269-273 (2023). https://doi.org/10.4174/astr.2023.104.5.269

18 Joh, H. C., Kim, M.-H., Ko, J. Y., Kim, J. S. & Jue, M. S. Evaluating the Performance of ChatGPT in a Dermatology Specialty Certificate Examination: A Comparative Analysis between English and Korean Language Settings. *Research Square* (2023). https://doi.org/https://doi.org/10.21203/rs.3.rs-3241164/v1

19 Koubaa, A. GPT-4 vs. GPT-3.5: A Concise Showdown. *TechRxiv* (2023). https://doi.org/10.36227/techrxiv.22312330.v2

20 Lim, Z. W. *et al.* Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine* **95**, 104770 (2023). https://doi.org/10.1016/j.ebiom.2023.104770

21 Raimondi, R. *et al.* Comparative analysis of large language models in the Royal College of Ophthalmologists fellowship exams. *Eye (Lond)* **37**, 3530-3533 (2023). https://doi.org/10.1038/s41433-023-02563-3

22 Ali, R. *et al.* Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery* (2023). https://doi.org/10.1227/neu.0000000000002551

23 Antaki, F., Touma, S., Milad, D., El-Khoury, J. & Duval, R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. *Ophthalmol Sci* **3**, 100324 (2023). https://doi.org/10.1016/j.xops.2023.100324