Evaluating the Performance of ChatGPT in Dermatology Specialty Certificate Examination-style Questions: A Comparative Analysis between English and Korean Language Settings

Hae C. Joh, Moon-Hwan Kim[1], Joo Y. Ko, Joung S. Kim[2], Mihn S. Jue

From the Department of Dermatology, College of Medicine, University of Hanyang, Seoul, [1]School of Electrical and Electronic Engineering, Yonsei University, Seoul, [2]Department of Dermatology, Hanyang University Guri Hospital, Hanyang University College of Medicine, Guri, South Korea
E-mail: zooooz@hanmail.net

Indian J Dermatol 2024:69(4):338-41

Dear Editor,

Artificial intelligence (AI) in the form of language models has gained substantial interest across various fields, including medicine.[1,2] One such model, Chat Generative Pre-trained Transformer (ChatGPT), uses deep learning algorithms to analyse and generate human-like responses using massive amounts of training data.[3] Previous research has related its performance to that of a third-year medical student in the United States Medical Licensing Examination, highlighting its potential role as a tutor in medical education.[4]

In dermatology, ChatGPT has various potential applications, such as producing clinical letters, leaflets, standardised reports, writing guidelines, and diagnosis and management planning.[5-8]

In Korea, the dermatology specialty certification examination (SCE) comprises 200 best-of-five multiple-choice questions (MCQs), half of which consists of clinical photographs and histological images and another half consists of text-based MCQs.[9]

Although ChatGPT shows promising prospects in medicine, the extent of its efficacy in dermatology requires further evaluation. Therefore, this study aimed to assess the performance of ChatGPT in the context of dermatology SCE-style questions consisting of text-based MCQs in Korea. Given that the primary language of the training data for ChatGPT is English, we hypothesised that the model's answers to identical MCQs may differ if the language used in the prompt is different.[10] Thus, this study also evaluated the performance variance of ChatGPT in the dermatology SCE-style questions when tasked with responding to the same questions in two different languages: English and Korean.

This study was conducted between 1 May and 23 June 2023. Three board-certified dermatologists (M.-S.-J., J.-Y.-K., and J.-S.-K.) collaborated to develop 200 text-based MCQs (four incorrect answers and one correct answer). These questions were designed based on the test blueprint of Dermatology SCE in Korea.

The questions were classified according to three cognitive functions: memorisation, judgment, and problem-solving. Afterward, they were sorted into evaluation categories, definition, cause, mechanism, symptoms and diagnosis, treatment and prognosis, and 10 dermatology topics.

Each question was fed into ChatGPT based on GPT-3.5, exactly once per session, and all responses were documented. The number of correct responses provided by the model in English and Korean was calculated, and qualitative analyses were performed to observe and record its reasoning process. To investigate whether variations in word order and phrasing influenced ChatGPT's responses, two additional sessions of asking questions were performed on the topics with the highest and lowest percentages of correct answers.

The correct response rates for both English and Korean were compared using univariable analysis (Chi-square test and Fisher's exact test). In addition, further analysis was performed for sub-classifications to compare the correct response rates between the two languages. All statistical analyses were conducted using IBM SPSS Statistics (IBM SPSS Statistics 25.0; IBM Corp., Armonk, NY, USA). Statistical significance was set at P value < 0.05.

The overall performance of ChatGPT during the examination is presented in Table 1.

In English, the overall correct response rate was 138/200 (69.0%), surpassing the 60% passing score set for the dermatology SCE in Korea. Further analyses were performed based on cognitive function, evaluation categories, and topics. Although the correct response rate was higher for 'memorisation' type questions at 62/86 (75.6%), no significant differences in performance were observed among the cognitive functions and evaluation categories (P > 0.05). Furthermore, within the evaluation categories, the correct response rates for the 'symptoms and diagnosis' sub-category were lower at 37/59 (62.7%), but the differences between evaluation categories were not statistically significant (P = 0.421). However, when evaluating the correct response rate according to topic, significant differences in performance were noted, with the sub-category of 'skin appendages and mucous membrane' obtaining the highest score of 16/18 (88.9%) and 'erythema, vascular disorders, urticarial' obtaining the lowest score of 3/11 (27.3%).

Table 1: Performance of ChatGPT in a Dermatology Specialty Certificate Examination-style Questions stratified according to evaluation category, cognitive function, and topic

Question Type	No. of Questions	No. of Correct Responses in English (%)	No. of Correct Responses in Korean (%)	*P English	*P Korea	**P- (English vs Korean)
All questions	200	138 (69.0)	114 (57.0)			0.013
Cognitive functions				0.249	0.022	
Memorisation	82	62 (75.6)	56 (68.3)			0.297
Judgment	54	35 (64.8)	25 (46.3)			0.053
Problem-solving	64	41 (64.0)	33 (51.6)			0.15
Evaluation categories				0.421	0.020	
Definition, Cause, Mechanism	71	52 (73.2)	47 (66.2)			0.361
Symptom, Diagnosis	59	37 (62.7)	25 (42.3)			0.221
Treatment, Prognosis	70	49 (70.0)	42 (60.0)			0.215
Topic				0.004	0.238	
Basic dermatology	18	15 (83.3)	14 (77.8)			0.105
Eczematous disorders	22	16 (72.7)	11 (50.0)			0.122
Papulosquamous disorders	19	12 (63.2)	8 (42.1)			0.194
Infectious disorders	29	24 (82.8)	18 (62.1)			0.078
Disorders of skin Appendages and mucous membranes	18	16 (88.9)	10 (55.6)			0.026
Autoimmune and vesiculobullous disorders	18	10 (55.6)	11 (61.1)			0.886
Pigmentation, metabolic, genetic, and endocrine disorders	18	9 (50.0)	6 (33.3)			0.310
Erythema, vascular disorders, urticaria	11	3 (27.3)	6 (54.5)			0.193
Benign and malignant neoplasms	18	15 (83.3)	10 (55.6)			0.070
Physical stimuli, fat atrophy, cosmetic dermatology, and others	29	18 (62.1)	20 (69.0)			0.454

*P-value – difference between subgroups, **P-value – difference between two languages

In the Korean language, the overall correct response rate was 114/200 (57.0%), which was slightly lower than the required passing score for the dermatology SCE in Korea. As in English, further analyses were performed based on cognitive function, evaluation category, and topic. The 'memorisation' type questions showed a higher correct response rate of 58/86 (68.4%), which proved to be statistically significant (P = 0.022). Furthermore, in the evaluation category, the 'symptoms and diagnosis' sub-category showed a significantly lower performance, with a correct response rate of 25/59 (42.3%) (P = 0.020). However, when evaluating the correct response rate according to topic, the highest score was observed in the 'Basic dermatology' sub-category at 14/18 (77.8%), while the 'Pigmentation, Metabolic, Genetic, and Endocrine Disorders' sub-category showed the lowest score at 6/18 (33.3%). However, these differences were not statistically significant.

A comparison between the English and Korean language settings revealed a statistically significant difference in the overall correct response rates (69.0% vs. 57.0%, P = 0.013), indicating a noteworthy variation in the performance of ChatGPT when utilising different languages.

However, when we compared the correct response rates between English and Korean languages, categorised according to cognitive function, evaluation category, and topic, we noticed a significant difference in performance between the two languages only in the sub-category of 'skin appendages and mucous membrane', with rates of 16/18 (88.9%) for English and 10/18 (55.6%) for Korean (P = 0.026).

While conducting qualitative analyses, we observed some inconsistencies in the performance of the ChatGPT model. Although the model was capable of answering questions correctly using proper reasoning, for some questions, the model gave correct responses but without appropriate reasoning. In addition, the model appeared to have difficulty in answering questions describing clinical situations that evaluated the cognitive function of judgment or problem-solving.

We added two more sessions of questions on topics that scored the highest and lowest percentages of correct answers in addition to the original session. While the difference in correct response rates between sessions was not statistically significant for either English or Korean, some notable changes were found from the prior answer, whether wrong or correct, and in the reasoning process used [Table 2].

In this study, ChatGPT achieved an overall correct response rate of 69.0% in a translated English version of the Korean Dermatology SCE-style questions, consistent with the results obtained by Passby *et al.*[11] in the UK

Table 2: Comparison of correct response rates over three sessions for topics in English and Korean. In each language, questions were asked in varying order of words and expressions across the sessions for the topic with the highest and lowest performance

Topic	No. of Questions	No. of Correct responses (%) (Session 1)	No. of Correct responses (%) (Session 2)	No. of Correct responses (%) (Session 3)	*P*
Questions in English					
Disorders of skin appendages and mucous membranes	18	16 (88.9)	16 (88.9)	17 (94.4)	0.802
Erythema, vascular disorders, urticaria	11	3 (27.3)	6 (54.6)	4 (36.3)	0.411
Questions in Korean					
Basic dermatology	18	14 (77.8)	15 (83.3)	14 (77.8)	0.892
Pigmentation, Metabolic, genetic, and endocrine disorders	18	6 (33.3)	8 (44.4)	8 (44.4)	0.736

P-value – difference in correct response rates across the sessions

Dermatology SCE using ChatGPT-3.5. However, when the same examination was performed using Korean, the correct response rate fell to 57.0%. This performance decline indicated a variation in the model's efficacy when a language other than English is used. We believed that because ChatGPT is trained using English-based texts, it demonstrates higher efficacy when answering questions presented in English. This discrepancy underlines the necessity of further fine-tuning and broadening the language capacity of such AI models to harness their potential optimally in a more global, multilingual context.[12]

When we sub-categorised the questions, we observed a significant difference in correct response rates within the cognitive function and evaluation categories in Korean and within topics in English. In the cognitive function domain, a higher correct response rate was observed for 'memorisation' type questions in both languages. This finding was consistent with those of a previous study by Bhayana *et al.*,[13] which showed superior performance of ChatGPT in answering questions requiring lower-order thinking skills.

Through a qualitative analysis of incorrect responses, we observed some inconsistencies in the performance of the model. This further emphasises the need for careful supervision and validation when employing AI tools, such as ChatGPT, in the dermatology field. Moreover, the language model seemed to have difficulty with questions describing clinical situations, reflecting a potential area of weakness in understanding or interpreting complex scenario-based clinical information that is essential for optimal patient care.

While our study highlighted the potential of AI language models in the dermatology field, it also underscored some of the model's current limitations. In the present study, we observed variations in the answers or reasoning processes for some questions when we asked the questions in varying order of words and expressions in further sessions for the topics with the highest and lowest correct response rates. Thus, this implies that LLMs do not remember specific facts or statements. Instead, their responses may vary depending on the new data they encounter. This highlights the unpredictable nature of current models and the need for continuous evaluation, particularly when they are used in the medical field.

Additionally, discrepancies between different languages pose inherent limitations to probabilistic inferences within language models. Likewise, ChatGPT did not perform well on higher-order thinking questions, which reflects a lack of training data specific to dermatology.

Nevertheless, as AI continues to evolve and mature, these limitations are anticipated to be addressed, allowing AI models such as ChatGPT to play an increasingly integral role in supporting medical fields worldwide.

The main limitation of our study is the exclusive focus on text-based MCQs. As a language model, ChatGPT-3.5 cannot directly analyse or interpret image contents. So we excluded questions associated with clinical photographs and histological images.

As AI continues to develop, it is expected to play an increasingly significant role in the medical field, including dermatology. However, its limitations must be acknowledged and addressed.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Li Z, Koban KC, Schenck TL, Giunta RE, Li Q, Sun Y. Artificial intelligence in dermatology image analysis:

Current developments and future trends. J Clin Med 2022;11:6826.
2. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023;388:1201-8.
3. Abdullah M, Madain A, Jararweh Y. ChatGPT: Fundamentals, applications and social impacts in Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), IEEE, 2022.
4. 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? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312.
5. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023;5:e179-81.
6. Manohar N, Prasad SS. Use of ChatGPT in academic publishing: A rare case of seronegative systemic lupus erythematosus in a patient with HIV infection. Cureus 2023;15:e34616.
7. Mondal H, Mondal S, Podder I. Using ChatGPT for writing articles for patients' education for dermatological diseases: A pilot study. Indian Dermatol Online J 2023;14:482-6.
8. Kluger N. Potential applications of ChatGPT in dermatology. J Eur Acad Dermatol Venereol 2023;37:e941-2.
9. Hwang I. Emerging tasks of specialty certifying examination: Educational measurement considerations. J Korean Med Assoc 2012;55:131-7.
10. Liu H, Ning R, Teng Z, Liu J, Zhou Q, Zhang Y. Evaluating the logical reasoning ability of ChatGPT and GPT-4. arXiv Preprint ArXiv:2304.03439 2023.
11. Passby L, Jenko N, Wernham A. Performance of ChatGPT on dermatology specialty certificate examination multiple choice questions. Clin Exp Dermatol 2023;llad197. doi: 10.1093/ced/llad197.
12. Lai VD, Ngo NT, Veyseh AP, Man H, Dernoncourt F, Bui T, *et al*. Chatgpt beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. arXiv Preprint ArXiv: 2304.05613 2023.
13. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: Insights into current strengths and limitations. Radiology 2023;307:e230582.

Access this article online

Quick Response Code:

Website: https://journals.lww.com/ijd

DOI: 10.4103/ijd.ijd_1050_23

How to cite this article: Joh HC, Kim MH, Ko JY, Kim JS, Jue MS. Evaluating the performance of ChatGPT in dermatology specialty certificate examination-style questions: A comparative analysis between English and Korean language settings. Indian J Dermatol 2024;69:338-41.

Received: November, 2023. **Accepted:** April, 2024.