Feasibility Study

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–8

DOI: 10.1177/20552076241236635
journals.sagepub.com/home/dhj

Sage

Performance of ChatGPT on the National Korean Occupational Therapy Licensing Examination

Si-An Lee[1], Seoyoon Heo[2] and Jin-Hyuck Park[3]

Abstract

Background: ChatGPT is an artificial intelligence-based large language model (LLM). ChatGPT has been widely applied in medicine, but its application in occupational therapy has been lacking.

Objective: This study examined the accuracy of ChatGPT on the National Korean Occupational Therapy Licensing Examination (NKOTLE) and investigated its potential for application in the field of occupational therapy.

Methods: ChatGPT 3.5 was used during the five years of the NKOTLE with Korean prompts. Multiple choice questions were entered manually by three dependent encoders, and scored according to the number of correct answers.

Results: During the most recent five years, ChatGPT did not achieve a passing score of 60% accuracy and exhibited interrater agreement of 0.6 or higher.

Conclusion: ChatGPT could not pass the NKOTLE but demonstrated a high level of agreement between raters. Even though the potential of ChatGPT to pass the NKOTLE is currently inadequate, it performed very close to the passing level even with only Korean prompts.

Keywords

ChatGPT, large language models, occupational therapy, licensing examination, artificial intelligence

Submission date: 25 August 2023; Acceptance date: 14 February 2024

Introduction

In recent decades, deep learning advancements have revolutionized artificial intelligence (AI) across various industries.[1–3] Notably, AI's ability to accurately classify traditional audio, images, and text data has enabled object categorization for photos and human-level text translation.[1–3] Recently, there has been significant interest in large language models (LLMs)-based AIs with unique capability to generate responses based on natural language input.

Unlike AIs limited to domain-specific data in a certain field, LLMs-based AIs can analyze non-domain-specific data. This characteristic eliminates the necessity of creating highly domain and problem-specific training data, thus enhancing their performances.[4] Given these advantages, the medical field has begun exploring the application of LLMs-based AIs for personalized healthcare, including diagnosis, clinical image analysis, and disease prediction.[5]

Consequently, there has been a surge of interest in leveraging Chatbot Generative Pre-trained Transformer (ChatGPT), a prominent LLM-based AI, within the

[1]Department of ICT convergence, The Graduate School, Soonchunhyang University, Asan, Republic of Korea
[2]Department of Occupational Therapy, Kyungbok University, Namyangju, Republic of Korea
[3]Department of Occupational Therapy, College of Medical Science, Soonchunhyang University, Asan, Republic of Korea

Corresponding author:
Jin-Hyuck Park, Department of Occupational Therapy, College of Medical Science, Soonchunhyang University, Room 1401, 22 Soonchunhyang-ro, Shinchang-myeon, Asan-si, Chungcheongnam-do 31538, Republic of Korea.
Email: roophy@naver.com

medical domain. ChatGPT is fine-tuned by LLMs that are trained based on text data from the Internet via reinforcement and supervised learning ways.[6] Several studies have examined ChatGPT's performance in medical licensing examinations to assess its ability to interact with patients based on medical knowledge. In a previous study, ChatGPT was employed in the United States Medical Licensing Examination (USMLE) and achieved a passing-level accuracy (60%) for medical-related natural language questions.[7] The clinical significance of ChatGPT lies in its ability to achieve a passing-level performance on the USMLE without needing for a professional human trainer. Furthermore, other studies demonstrated that ChatGPT can also pass the Japanese Medical Licensing Examination (JMLE) and Taiwan pharmacist licensing examination.[8,9] In other words, ChatGPT could serve as a tool for medical assistance or self-study for medical students across different countries,[10] even in non-English-language studies where prompts were presented in both native languages and English translations.[8,9]

Although there have been some studies on ChatGPT for pharmacist and nurse licensing examinations,[9,11] the performance of ChatGPT has been primarily evaluated in the context of medical licensing examinations. Therefore, its applicability in other medical fields remains uncertain. Physicians and occupational therapists (OTs) both play crucial roles in the rehabilitation system. However, their specialties and responsibilities differ. Physicians primarily focus on diagnosing and treating illness, while OTs are heavily involved in facilitating clients' engagement in occupations. Specifically, OTs address various aspects of clients' performances such as physical, cognitive, psychological, and sensory-perceptual factors to support their engagement in occupations.[12]

Given the diversity of medical fields, this study aimed to assess the performance of ChatGPT using questions from the National Korean Occupational Therapy Licensing Examination (NKOTLE). The NKOTLE encompasses all knowledge domains essential for occupational therapists.[13] The difficulty and complexity of the NKOTLE questions are highly standardized and regulated by a panel of experts, making them suitable for AI testing. These questions are well-established, with very stable raw scores and psychometric properties over the past five years.[14] In addition, as the NKOTLE questions are exclusively in a multiple-choice, and text-oriented format, they provide a challenging assessment for ChatGPT. However, no previous studies have reported the performance of ChatGPT on the NKOTLE.

Therefore, this study aimed to determine if ChatGPT could successfully pass the last five years of the NKOTLE, offering quantitative feedback on its performance and evaluating its potential for application in the field of occupational therapy, similar to other medical domains. Additionally, unlike prior works that utilized prompts in both native language and English translations, this study sought to verify the feasibility of using a non-English language-based prompt by presenting all questions exclusively in Korean.

Methods

Artificial intelligence

ChatGPT is an advanced language model developed by OpenAI, located in San Francisco, CA, USA. This model utilizes self-attention mechanisms and extensive training data to generate coherent and contextually appropriate responses in natural language within a conversational context. It excels in handling long-range dependencies, ensuring that its generated responses are well-connected and relevant. Unlike other conversational systems or Chatbots that have access to external sources of information, such as internet searches or databases, ChatGPT is a self-contained server-based model. Consequently, all responses it produces are generated within the model itself, based on abstract relationships between words or "tokens" within its neural network.[7] In this study, the freely available version 3.5 of ChatGPT was utilized.

Input source

Public-available test questions from NKOTLE-2018 to NKOTLE-2022 were from the official NKOTLE website. Due to the lack of access to previous data, only questions from the past five years were utilized for this study. The NKOTLE is taken by students who are planning to graduate from a three- or four-year occupational therapy program in South Korea. The NKOTLE consists of three units. The first unit has two sub-units. The first sub-unit assesses basic knowledge of occupational therapy including anatomy, physiology, and public health with 70 questions, and the second sub-unit tests the Medical Service Act with 20 questions. The second unit tests specialized occupational therapy knowledge such as neurological, musculoskeletal, and psychiatric occupational therapy with 100 questions. The third unit assesses clinical reasoning through a 50-question paper-and-pencil practical examination that provides illustrations and hypothetical clinical data. The passing threshold is 60% or more correct answers for each unit, and less than 40% of correct answers for each unit will fail regardless of the overall percentage of correct answers. Students who pass this examination are licensed by the Ministry of Health and Welfare of South Korea and registered as occupational therapists.[13]

In this study, only the first and second units of the NKOTLE problems were utilized due to copyright restrictions preventing access to the third unit, which is not disclosed on the NKOTLE website. Furthermore, questions of the third unit primarily contain images or graphs that depict medical conditions of virtual cases, making it unsuitable for prompt input in

the CHATGPT 3.5 version. Both the first and second units consist entirely of multiple-choice questions with a single answer, where forced justification is not required.

To ensure that the input data used in the study were representative, 20 test questions per year were randomly sampled from the first unit and second unit questions. It was verified that these NKOTLE questions were not indexed in Google after January 1, 2018. Subsequently, it was confirmed that these questions from the first and second units did not include images or graphs. After filtering, a total of 950 questions (190 questions per year) were advanced to encoding.

Encoding

This study encoded by reproducing the original NKOTLE questions verbatim. To ensure a single correct answer, the following prompt was consistently encoded with each question: "Which of the following best represents the most appropriate answer?" Three encoders independently encoded the prompt. A new chat session was started in ChatGPT for each entry to reduce memory retention bias. If ChatGPT failed to provide an answer for a question in terms of the answer choice number or text, the question was re-encoded up to three times.

Accuracy and interrater agreement

The three encoders independently evaluated the accuracy of ChatGPT based on the criterion for determining the correct answer. According to this criterion, ChatGPT should either present the correct answer number or the text corresponding to the correct number as provided in the NKTOLE. The correct answer for each question was coded in an Excel file and shared with an independent examiner. The independent examiner analyzed the percentage of correct answers determined by encoders, compared it against the passing criteria, and assessed the inter-rater agreement between encoders. A schematic of the study flow is provided in Figure 1.

Statistical analysis

All data were analyzed using IBM SPSS Statistics version 22.0. The Fleiss' kappa statistics were used to compute inter-rater agreement on the accuracy of ChatGPT on the NKOTLE. The Spearman correlation test was conducted to confirm the correlation between accuracy and inter-rater agreement. Statistical significance was set at $p < 0.05$.

Results

Accuracy

The performance (correct, incorrect, and undetermined answers) of ChatGPT on the NKOTLE by year is presented in Figure 2. In the 2018 NKOTLE, the average accuracy of the three units was 52.2%. In the 2019 NKOTLE, the

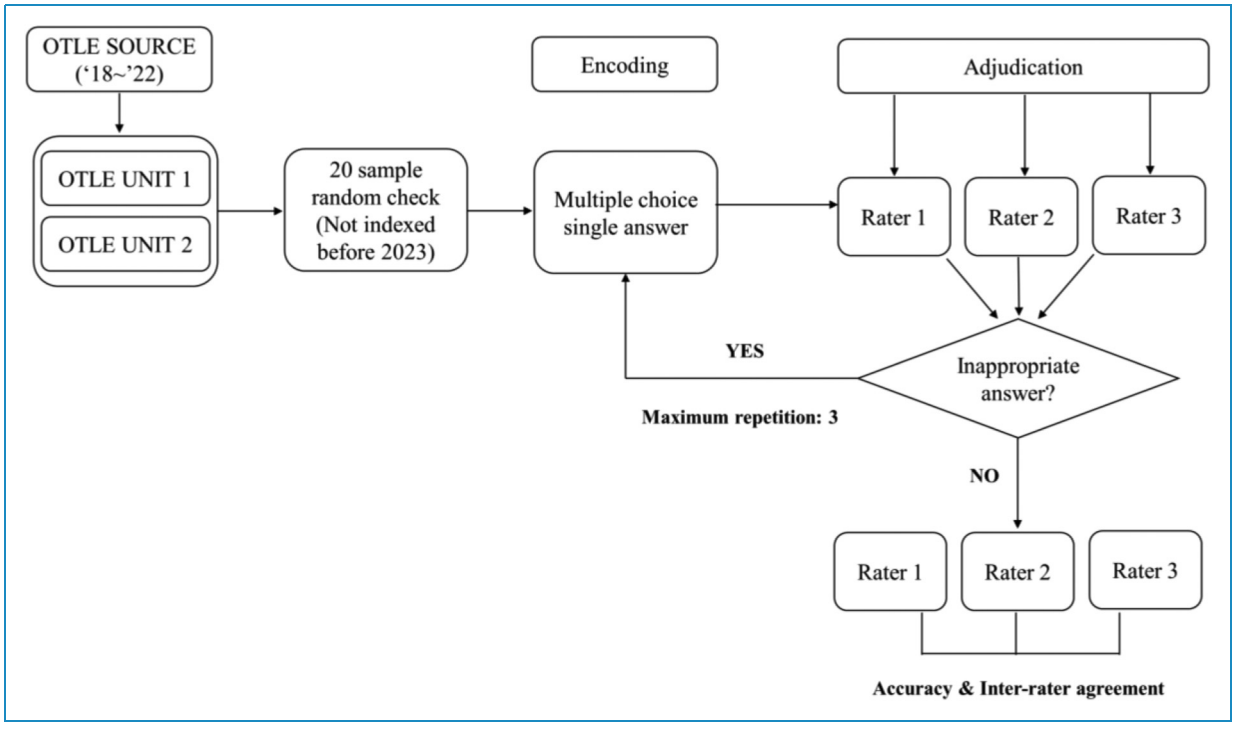

Figure 1. Flowchart for sourcing, encoding, and adjudicating results.

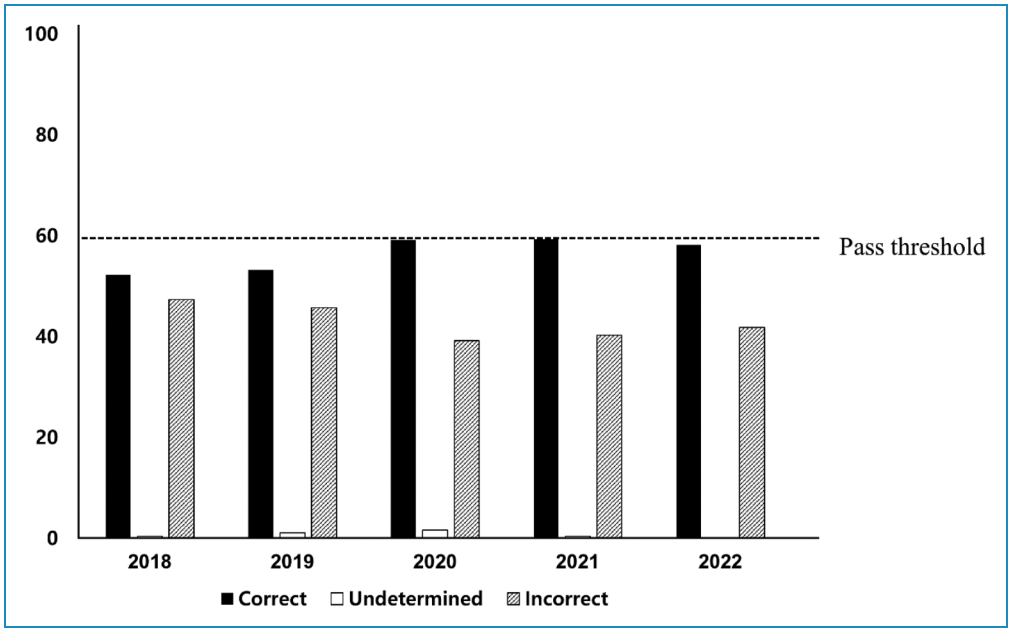

Figure 2. Accuracy of ChatGPT on the NKOTLE.

average accuracy of the three units was 53.2%. In the 2020 NKOTLE, the average accuracy of the three units was 59.2%. In the 2021 NKOTLE, the average accuracy of the three units was 59.3%. In the 2022 NKOTLE, the average accuracy of the three units was 58.2%. For both 2020 and 2021 NKOTLEs, ChatGPT showed a near-passing accuracy. Undetermined answers were counted as incorrect and then the accuracy of ChatGPT was calculated. Table 1 shows each rater's ChatGPT accuracy across the past five years of the NKOTLE.

Interrater agreement

Regarding the inter-rater reliability of ChatGPT between raters, the Fleiss kappa ranged from 0.602 to 0.845 ($p's < 0.01$) (Table 1), suggesting that the inter-rater reliability of ChatGPT on the NKOTLE was acceptable. This finding means that ChatGPT's performance on the KNOTLE does not depends on raters.

Relationship between accuracy and interrater agreement

Spearman correlation revealed that there was no significant correlation between the accuracy of ChatGPT on the NKOTLE and inter-rater agreement ($r = .420$, $p = .119$). This finding suggested that the accuracy of ChatGPT was not mediated by inter-rater agreement. In other words, ChatGPT's performance on the NKOTLE solely depends on its ability to select the correct answer.

Discussion

Our study aimed to examine the feasibility of ChatGPT on the NKOTLE. The current findings demonstrated that ChatGPT could not pass the examination. Nevertheless, ChatGPT was close to the passing score with only a few more correct answers in some years of the NKOTLE. To sum up, the proficiency and ability of ChatGPT to interpret questions pertaining to the NKOTLE are currently inadequate.

ChatGPT emergence marks a significant advancement in natural language processing.[15] Its continuous growth promises profound impacts across various sectors such as business, medicine, education, and entertainment.[16] By leveraging its extensive language data learning, ChatGPT crafts human-like text and can revolutionize education.[15] Particularly in medical studies, it offers personalized learning, aids in exam inquiries, and enhances student engagement.[17] In this vein, while ChatGPT has been evaluated for exams in a variety of healthcare fields, to the best of our knowledge, this is the first study to determine the performance of ChatGPT on an occupational therapy licensing examination.

In this study, across the five years of the NKOTLE, ChatGPT's score did not meet the passing requirement. According to statistics, the average pass rate of the NKOTLE across the five years was 88.8%.[18] Compared to occupational therapy students who have undergone traditional 3 or 4-year education, ChatGPT's performance is currently insufficient. Considering that ChatGPT 3.5 works well with multiple-choice questions and that NKOTLE questions

Table 1. ChatGPT accuracy across the past five years of the NKOTLE.

		Accuracy (%)					
NKOTLE		First rater	Second rater	Third rater	Avg	Benchmark (>60.0)	Fleiss kappa
2018	Unit 1-1	67.1	70.0	67.1	68.1	Fail (59.2)	0.744***
	Unit 1-2	25.0	20.0	25.0	23.3		0.756***
	Unit 2	73.0	63.0	60.0	65.3		0.719***
2019	Unit 1-1	68.6	64.3	65.7	66.2	Fail (53.2)	0.833***
	Unit 1-2	40.0	35.0	35.0	36.7		0.656**
	Unit 2	51.0	58.0	61.0	56.7		0.741***
2020	Unit 1-1	78.6	65.7	67.1	70.5	Fail (59.2)	0.845***
	Unit 1-2	45.0	50.0	50.0	48.3		0.831***
	Unit 2	61.0	57.0	58.0	58.7		0.813***
2021	Unit 1-1	65.7	60.0	57.1	61.0	Fail (59.3)	0.672***
	Unit 1-2	75.0	40.0	50.0	55.0		0.602**
	Unit 2	58.0	61.0	67.0	62.0		0.751***
2022	Unit 1-1	54.3	71.4	70.0	65.2	Fail (58.2)	0.807***
	Unit 1-2	50.0	35.0	60.0	48.3		0.660**
	Unit 2	68.0	56.0	59.0	61.0		0.813***

**$p<0.01$, ***$p<0.001$.

used in this study are all multiple-choice questions, a high accuracy of ChatGPT was expected. However, ChatGPT did not pass the 2018–2022 examinations. This misalignment with the expectation might be due to the following reasons. Firstly, there are differences in medical laws and policies between South Korea and the States. Specifically, when comparing accuracy by units, it was found that the accuracy of UNIT 1–2 was consistently lower than that of other units. This was because UNIT 1–2 consisted of questions related to medical laws and policies in South Korea, which ChatGPT might lack training on.[19] This is in contrast to prior studies showing that ChatGPT came close to a passing score in the English exam.[9] Indeed, English is the most resource-rich language in many applications of natural language processing.[20] In other words, ChatGPT performs better in English in other languages,[21] supporting our assumptions. Consequently, ChatGPT performed relatively well in UNIT 1–1 and UNIT 2 as these units cover subjects such as anatomy, physiology, and specialized occupational therapy areas, which are less than affected by language or national conditions. The accuracy for these units is similar to that of ChatGPT applied to licensing examinations in English or English-speaking cultures.[7–9,11] In addition to the constrained accuracy stemming from limitations in English-related learning data, English prompts themselves confer distinct advantages. In prior studies, ChatGPT 3.5 yielded slightly higher accuracy within the range of less than 5% when prompts were entered in English as opposed to native languages.[8,9] This was particularly noticeable in questions related to the administration and law of non-English speaking countries, domains presumed to have limited training data with ChatGPT 3.5. These findings underscore the superiority not only of cultural consideration but also of utilizing English prompts in enhancing ChatGPT 3.5's performance.[9] Nevertheless, when contemplating the substantial variance in accuracy between culturally influenced and non-cultural items, it becomes evident that distinctions in training data exert a more pronounced influence on

accuracy compared to the impact of prompts. Secondly, all questions in the NKOTLE required the selection of the best answer, while some questions had multiple-choice answers provided by ChatGPT, which could be regarded as suboptimal rather than incorrect in clinical practice,[19] which is consistent with a previous study.[19] Indeed, in a previous study, since ChatGPT sometimes did not understand multiple-choice questions, ChatGPT selected 2 or more answers rather than choosing the best answer.[19] These findings suggest that ChatGPT is not yet trained to choose only the best option.

On the other hand, across the five years of the NKOTLE, a high agreement between three raters was observed, indicating that ChatGPT could provide reliable outputs regardless of who would use it. Given that ChatGPT can produce different outputs depending on prompts, this study used uniform prompts. This suggests that using controlled prompts can produce reliable results, supported by a previous study.[8,11,22] A previous study showed that inter-rater agreement is positively correlated with accuracy.[7] Therefore, it was assumed that the low accuracy was heavily due to missing information rather than overcommitment to incorrect answer choices. In contrast, in this study, there was no significant correlation between accuracy and inter-rater agreement. Thus, the low accuracy of ChatGPT on the NKOTLE could be attributed to incorrect answer choices rather than being based on inconsistent answer choices between raters, which is different from the findings of a previous study.[7] In other words, the current ChatGPT's performance was assessed by the consistent answers given by ChatGPT.

Justifications for correct answers were not investigated as they did not require justification. However, given the low accuracy of ChatGPT on the NKOTLE, ChatGPT is not sufficiently robust at present to be utilized for the educational demands of occupational therapy students. However, the findings of this study should not be generalized to other subjects or fields as ChatGPT's knowledge will rapidly improve in response to user feedback, leading to different outcomes of subsequent trials using the same questions.[8,11,16] Therefore, future studies should investigate whether ChatGPT can also be used for educational purposes to support the process of writing explanations for occupational therapy-related questions. This will help occupational therapy students acquire knowledge in their specialty by using ChatGPT as a learning tool without the need to use other learning media.[7]

While most of the previous studies reported ChatGPT's performance using only English or both native language and English translation as a prompt, this study confirmed similar levels of accuracy to the previous studies,[8,9,11,23] even though only Korean prompts were used. This can be attributed to the fact that most prior studies were conducted in regions outside the English-speaking culture, specifically applying ChatGPT to licensing examinations for doctors, nurses, and pharmacists in Japan, China, and Taiwan.[8,9,11,23] Meanwhile, even ChatGPT's performance on questions within English culture reveals a similar accuracy level compared to the current findings.[7] This suggests that ChatGPT 3.5 retains multilingual characteristics despite the predominantly English-centric training data.[4,22] The current findings are especially interesting considering that the training data of LLMs are centered on English,[4,24] suggesting that even in non-English-speaking countries, ChatGPT could be utilized in its native language as long as questions are free from language and national conditions. Nevertheless, the impact of English as a prompt could not be ignored as a previous study reported the superiority of English-translated prompts over native-language prompts. This might be overcome to some extent in the future by using extensions such as automatic English translations.[25] Another issue revolves around potential biases in the AI model, primarily stemming from ChatGPT's training data and user prompts. To address these potential biases, future studies involve scrutinizing the composition of the training dataset and evaluating the effectiveness of any bias mitigation techniques applied. Additionally, the prompts will be reviewed to ensure neutrality and inclusivity. We have added this perspective to the discussion part.

This study has some limitations. Firstly, we assumed that this study applied ChatGPT to the NKOTLE for the first time. However, this could not be verified. Secondly, due to the lack of disclosure, this study was unable to verify ChatGPT's performance on UNIT 3. Thus, it was not possible to confirm the final KNOTLE pass. Nevertheless, it is currently impossible to verify ChatGPT 3.5 because UNIT 3 contains several images, figures, or tables, which are not interpreted by ChatGPT 3.5. Finally, to improve ChatGPT's performance, this study used a simple prompting method instead of using the original question as a prompt. However, the accuracy might be lower due to the lack of optimized prompt engineering.[8] Nonetheless, the results of this study are highly reproducible as this study confirmed ChatGPT performance with an almost exact replication of the actual question format. However, in the future, it is necessary to prompt ChatGPT so that it does not give incorrect answers when it is unsure of the answer.

In conclusion, ChatGPT's knowledge in answering the NKOTLE is not yet comparable to that of occupational therapy students in Korea. However, its ability continues to evolve. We expect the higher performance of ChatGPT with prompt engineering and English translation prompts in the future. Therefore, in the near future, professors and students of occupational therapy need to pay attention to the potential of AI chatbots and consider their applications in learning and teaching methods. Specifically, since ChatGPT provides correct answers to each test question along with the theoretical background of the answer, we expect that it will be actively used for medical education by utilizing the advantage that users can learn without

having to review and search additional literature besides ChatGPT.

Acknowledgements: We would like to thank undergraduate students of department of occupational therapy from Soonchunhyang university for their assistance.

Contributorship: S-AL and J-HP contributed to the design of the study. S-AL, SH, and J-HP contributed to the collection and analysis of the data. SH and J-HP contributed to the draft of the article. All authors reviewed and edited the manuscript and approved the final version of the manuscript

Data availability statement: The authors confirm that the data supporting the findings of this study are available from the corresponding author on request.

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Soonchunhyang University Research Fund. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by Ministry of Education (no. 2021R1I1A3041487).

Ethical approval: Not applicable.

Consent statement: Not applicable as this study is neither a clinical trial nor a human trial.

Guarantor: J-HP.

ORCID iDs: Si-An Lee https://orcid.org/0009-0004-5597-9778
Jin-Hyuck Park https://orcid.org/0000-0002-0222-8901

References

1. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016, pp.2818–2826.
2. Zhang W, Feng Y, Meng F, et al. Bridging the gap between training and inference for neural machine translation. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 28 July–2 Aug 2019, pp.4334–4343.
3. Wang W and Siau K. Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: a review and research agenda. *J Database Manag* 2019; 30: 61–79.
4. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst* 2020; 33: 1877–1901.
5. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* 2023; 11: 887.
6. Castelvecchi D. *Are ChatGPT and AlphaCode going to replace programmers?*, https://www.nature.com/articles/d41586-022-04383-z (2022, accessed 1 August 2023).
7. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health* 2023; 2: e0000198.
8. Kasai J, Kasai Y, Sakaguchi K, et al. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations. *arXiv [csCL]*, https://arxiv.org/abs/2303.18027 (2023).
9. Wang YM, Shen HW and Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. *J Chin Med Assoc* 2023; 86: 653–658.
10. Nisar S and Aslam MS. Is ChatGPT a good tool for T&CM students in studying pharmacology?, https://doi.org/10.2139/ssrn.4324310 (2023, accessed 1 August 2023).
11. Taira K, Itaya T and Hanada A. Performance of the large language model ChatGPT on the national nurse examinations in Japan: evaluation study. *JMIR Nurs* 2023; 6: e47305.
12. Boop C, Cahill SM, Davis C, et al. Occupational therapy practice framework: domain and process fourth edition. *Am J Occup Ther* 2020; 74: 1–84.
13. Information on the National Korean Occupational Therapy Licensing Examination. Korea Health Personnel Licensing Examination Institute, https://www.kuksiwon.or.kr/subcnt/c_2015/1/view.do?seq=7&itm_seq=13 (accessed 1 August 2023).
14. Performance Data. Korea Health Personnel Licensing Examination Institute, https://www.kuksiwon.or.kr/peryearPass/list.do?seq=13&srchWord=13 (accessed 1 August 2023).
15. Biswas S. ChatGPT and the future of medical writing. *Radiology* 2023; 307: e223312.
16. Hacker P, Engel A and Mauer M. Regulating ChatGPT and other large generative AI models. *arXiv [csCL]*, https://arxiv.org/abs/2302.02337 (2023).
17. Jeblick K, Schachtner B, Dexl J, et al. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv [csCL]*, https://arxiv.org/abs/2212.14882 (2022).
18. Performance Data. Korea Health Personnel Licensing Examination Institute, https://www.kuksiwon.or.kr/news/brd/m_54/view.do?seq=453&&itm_seq_1=0&&itm_seq_2=0 (accessed 1 August 2023).
19. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof* 2023; 20: 1516081869.
20. Névéol A, Dalianis H, Velupillai S, et al. Clinical natural language processing in languages other than English: opportunities and challenges. *J Biomed Semantics* 2018; 9: 1–13.
21. Hu J, Ruder S, Siddhant A, et al. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In: *Proceedings of the 39th International*

Conference on Machine Learning, Baltimore, Maryland, USA, 17–23 July 2022, pp.4411–4421.
22. Liu Y, Gu J, Goyal N, et al. Multilingual denoising pre-training for neural machine translation. *Trans Assoc Comput Linguist* 2020; 8: 726–742.
23. Wang X, Gong Z, Wang G, et al. ChatGPT performs on the Chinese national medical licensing examination. *J Med Syst* 2023; 47: 86.
24. Zhang S, Roller S, Goyal N, et al. Opt: Open pre-trained transformer language models. *arXiv [csCL]*, https://arxiv.org/abs/2205.01068 (2022).
25. Fang C, et al. How does ChatGPT4 perform on Non-English National Medical Licensing Examination? An Evaluation in Chinese Language. *medRxiv*, https://www.medrxiv.org/content/10.1101/2023.05.03.23289443v1 (2023).