Brief report

Is ChatGPT's knowledge and interpretation ability comparable to that of medical students in Korea for taking a parasitology examination?: a descriptive study"

Sun Huh

Department of Parasitology and Institute of Medical Education, College of Medicine, Hallym University, Chuncheon, Korea

*Corresponding email: shuh@hallym.ac.kr

Word count of abstract:

Word count of text: 1412

No. of tables: 2

No. of figure: 1

Abstract

It aims to compare the knowledge and interpretation skills of ChatGPT, a language model of artificial general intelligence, with those of medical students in Korea. This was done by administering a parasitology examination to both ChatGPT and the medical students. The examination consisted of 79 items and was administered on January 1, 2023. The results of the examination were analyzed in terms of ChatGPT's overall performance score, its correct answer rate by knowledge level of the items, and the acceptability of its explanations of the items. The results showed that ChatGPT's performance was lower than that of the medical students and that ChatGPT's correct answer rate was not related to the knowledge level of the items. However, there

was a relationship between an acceptable explanation and a correct answer. In conclusion, ChatGPT's knowledge and interpretation abilities on the parasitology examination are not yet comparable to those of medical students in Korea.
Keywords: Artificial intelligence; Educational measurement; Knowledge; Medical students; Republic of Korea

Background: Siobhan O'Connor [1] wrote an editorial with the opening paragraphs written by ChatGPT, an artificial intelligence (AI) chatbot. ChatGTP was trained by the model using Reinforcement Learning from Human Feedback, using the same methods as InstructGPT (GPT: Generative Pre-trained Transformer) [2]. ChatGPT argued that AI chatbots could be used to provide tutoring and homework help by answering questions and providing explanations to help students understand difficult concepts. However, there are concerns that the use of AI software by nursing students to write university assessments could lead to a decrease in the value of the assessments and the overall quality of the nursing program. After the release of ChatGPT to the public on November 30, 2022, it became a hot topic, particularly in the field of education. Chris Stokel-Walker also noted that ChatGPT, an AI-powered chatbot that generates intelligent-sounding text in response to user prompts, including homework assignments and exam-style questions, has caused concern [3]. It is critical for medical students to be able to evaluate the accuracy of medical information generated by AI and to have the skills to create reliable, validated information for patients and the public. Therefore, it is necessary to determine how accurately ChatGPT, a recently developed AI chatbot, can solve questions on a medical examination. This comparison of ChatGPT's abilities may provide an incentive for medical students to use ChatGPT for their learning.

Objectives: The aim of this study is to compare the knowledge and interpretation skills of ChatGPT with those of medical students in Korea by administering a parasitology examination, a subject that is essential in medical schools in Korea. Specifically, the following will be investigated:

1) the scores of ChatGPT compared to those of the medical students; 2) the correct answer rate of ChatGPT also compared by the knowledge level of the items; and 3) the acceptability of ChatGPT's explanations as current parasitology knowledge, as evaluated by the author.

Ethics statement: This is not a study of human subjects but an analysis of the results of an educational examination routinely conducted in medical college. Therefore, neither approval by the institutional review board nor obtaining informed consent is required.

Study design: This is a descriptive study to compare the ability of ChatGPT to solve questions with that of medical students.

Setting: On January 1, 2022 (Seoul time), the same items on a parasitology examination administered to first-year medical students at Hallym University on December 12, 2022, on a computer-based testing (Supplement 1) were used for ChatGPT (version Dec 15, 2022). The answers given by ChatGPT were compared to those of the medical students.

Participants: A total of 77 medical students took the parasitology exam in the fourth quarter of 2022. ChatGPT was counted as one examinee. There were no exclusion criteria.

Variables: The knowledge level of the items and the scores of the examinees were the variables.

Data sources and measurement: The response data of 77 medical students on the parasitology examination and ChatGPT were compared. The explanation data provided by ChatGPT (Supplement 2, Fig. 1) were also evaluated for their acceptability by the author. The comparison of correct answers for items according to the level of knowledge was analyzed. Acceptability of explanation was classified as good, good but needing revision, and not acceptable.

Bias: There was no bias in the selection of examinees. All students who attended the parasitology lecture were included.

Study size: Sample size estimation was not required because all target students were included and one AI platform was added.

Statistical methods: Descriptive statistics were used to analyze the score of the Chatbot. Comparison analysis was done using DBSTAT version 5.0

Score by ChatGPT and Comparison with the Medical Students' Performance

According to data from Dataset 1, ChatGPT correctly answered 48 out of 79 total items (60.8%). This score is lower than the average score of 77 medical students, which was 70.8 out of 79 (89.6%), with a minimum score of 63 (79.7%) among the medical students.

Comparison of Correct Answer Rate by ChatGPT According to Knowledge Level of Items

Table 1 shows the responses of ChatGPT according to the knowledge level of the items. The chi-square test showed that the statistic was $x^2 = 3.02$, df = 2, with a significance level of 0.05 ($x^2 = 5.99$). The result indicates that the relationship between the two variables is not significant (P = 0.2206).

Table 1. Correct Responses by ChatGPT According to the Knowledge Level of 79 Items

Knowledge Level of Items	Correct Response	Incorrect Answer
Recall	17	15
Interpretation	20	12
Problem-Solving	11	4

Acceptability of ChatGPT's Explanation

Table 2 shows the acceptability of ChatGPT's explanations according to the correctness of the answer. The chi-square test showed that the statistic was $x^2 = 51.62$, df = 2, with a significance level of 0.05 ($x^2 = 5.99$). The result indicates that the relationship between the two variables is significant (P = 0.0000).

Table 2. Acceptability of ChatGPT's Explanation on the 79 Question Items by Correctness of Answer

Explanation	Correct Answer	Incorrect Answer
Good	41	3
Need to be revised	7	8
Not acceptable	0	20

Key results: ChatGPT's performance was lower than that of medical students. The correct answer rate for ChatGPT was not related to the knowledge level of the items. There was an association between an acceptable explanation and a correct answer.

Interpretation: ChatGPT's correct answer rate of 60.8% was not a poor performance, as the questions were not easy for medical students to answer correctly. The high average score (89.6%) of the medical students may be due to their prior learning of parasitology and the examination being taken immediately after the class. If the examination were taken one or two months after the class, the students' performance scores may be lower. Some incorrect answers may be due to the

following factors: first, ChatGPT is currently unable to interpret figures, graphs, and tables as a student can, so the author had to describe these materials in text form. Second, some epidemiological data unique to Korea were outside of ChatGPT's knowledge. Some of this data is written in Korean or not searchable online. Third, ChatGPT sometimes did not understand multiple choice questions where the examinee must select the best answer out of multiple options. ChatGPT sometimes selected two or more options, as it has not yet been trained to do otherwise.

There was no difference in correct answers according to the knowledge level of the items. However, this may vary in different examinations and may be a unique phenomenon for this parasitology exam. ChatGPT's explanations for the question items were generally acceptable if it made a correct selection, but the explanation for 7 items needed to be updated or revised as they contained incorrect information. This suggests that ChatGPT's knowledge in specific fields, for example parasitology is still not sufficient. If the correct option was selected, the explanation was not acceptable or needed revision in 90.0% of items. This is an anticipated result, as explanations for incorrect selections by students are usually not acceptable.

Comparison with previous studies: There have been no reported studies on the comparability of ChatGPT's performance on medical examinations.

Limitations: The input for the question items for ChatGPT was not exactly the same as for the medical students. Chatbot is unable to receive information in the form of graphs, figures, and tables, so this information was re-described by the author. Additionally, interpretation of the explanations and correct answers may vary according to the perspective of different parasitologists, although the author has worked in the field of parasitology for 40 years (1982-2022) in Korea. Best practices for patient care may also vary according to region and medical environment.

Generalizability: The above results cannot be generalized to other subjects or medical schools directly, as Chatbot will likely continue to evolve rapidly through feedback from users. A future trial

with the same items may yield different results. The present results reflect the abilities of ChatGPT on January 1, 2023.

Implications for medical/health students and professors to use ChatGPT: Currently, ChatGPT's level of knowledge and interpretation is not sufficient to be used by medical students, especially in medical school exams. This may also be the case for high-stakes exams, including health licensing exams. However, I believe that ChatGPT's knowledge and interpretation skills will improve rapidly through deep learning, similar to AlphaGo's ability [5]. Therefore, medical/health professors and students should be mindful of how to incorporate this AI platform into medical/health education in the near future. Furthermore, AI should be integrated into the medical school curriculum, although some schools have already adopted it [6].

Conclusion: ChatGPT's knowledge and interpretation ability in answering the parasitology examination is not yet comparable to that of medical students in Korea. However, these abilities will likely improve through deep learning. Medical/health professors and students should be aware of the progress of this AI chatbot and consider its potential adoption in learning and education.

ORCID

Sun Huh: https://orcid.org/0000-0002-8559-8640

Conflict of interest

There is no conflict of interest to be reported.

Author's contribution

All work was done by Sun Huh.

Funding

None

Data availability

Dataset 1. Raw data for analysis, including item number, knowledge level, correct answer, ChatGPT's answer, and correctness of exaplanation for parasiology examination taken by the first year medical students, Hallym University on December 12, 2022.

Acknowledgments

None

Supplementary materials

Supplement 1. Seventy nine items of parasitology examination taken by the first year medical students, Hallym University on December 12, 2022

Supplement 2. ChatGPT's response to 79 items of parasitology examination taken by the first year medical students, Hallym University on December 12, 2022, inputted on January 1, 2023 by the author. Figures and tables are removed and explained in the item stem. Explanation on the selecting option by ChatGPT was also included.

References

1. Stokel-Walker C. AI bot ChatGPT writes smart essays - should professors worry? Nature 2022 Dec 9. https://doi.org/10.1038/d41586-022-04397-7

2. OpenAI. ChatGPT Dec 15 version [Internet]. 2022 [cited 2023 Jan 1]. Available from: https://chat.openai.com/chat

3. O'Connor S, ChatGPT. Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? Nurse Educ Pract 2022;66:103537. https://doi.org/10.1016/j.nepr.2022.103537

4. Park SH, Do KH, Kim S, Park JH, Lim YS. What should medical students know about artificial intelligence in medicine? J Educ Eval Health Prof 2019;16:18.
https://doi.org/10.3352/jeehp.2019.16.18
5. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D. Mastering the game of Go with deep neural networks and tree search. Nature. 2016;529:484-489.
https://doi.org/10.1038/nature16961
6. Hu R, Fan KY, Pandey P, Hu Z, Yau O, Teng M, Wang P, Li A, Ashraf M, Singla R. Insights from teaching artificial intelligence to medical students in Canada. Commun Med (Lond). 2022;2:63.
https://doi.org/10.1038/s43856-022-00125-4

Explanation for figures

Fig. 1. Screenshot of ChatGPT's answer to the question item of parasitology examination for medical students in Hallym University.